A HEDONIC APPROACH TO ESTIMATING SOFTWARE COST USING ORDINARY LEAST SQUARES REGRESSION AND NOMINAL ATTRIBUTE VARIABLES

THESIS

Marc D. Ellis, Captain, USAF

AFIT/GCA/ENV/06M-04

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GCA/ENV/06M-04

A HEDONIC APPROACH TO ESTIMATING SOFTWARE COST USING ORDINARY LEAST SQUARES REGRESSION AND NOMINAL ATTRIBUTE VARIABLES

THESIS

Presented to the Faculty

Department of Systems and Engineering Management

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Cost Analysis

Marc D. Ellis, BS

Captain, USAF

March 2006

AFIT/GCA/ENV/06M-04

A HEDONIC APPROACH TO ESTIMATING SOFTWARE COST

USING ORDINARY LEAST SQUARES REGRESSION

AND NOMINAL ATTRIBUTE VARIABLES

Marc D. Ellis, BS
Captain, USAF

Approved:

/signed/	17 March 2006
Michael J. Hicks, Ph.D. (Chairman)	date
/signed/	17 March 2006
Curtis G. Tenney, Lt Col, USAF, Ph.D. (Member)	date
/signed/	17 March 2006
Jeffrey S. Smith, Maj, USAF, Ph.D. (Member)	date

Abstract

Software spending is increasing within the DoD, NASA, and other technologically advanced organizations, with significant effects on program budgets. Cost estimators must have the best tools available. However, many current models are problematic due to inaccuracy and unavailability of the input parameters, the technical expertise and expense required to operate them, and the difficulty in explaining their outputs. Two databases were analyzed: 60 NASA/JPL software projects and 116 projects from the International Software Benchmarking and Standards Group database. Models developed using ordinary least squares regression employed parameters representing the presence of project characteristics. The models' predictive characteristics were compared based on the source database, project size, model transformations, and the variable combinations. COCOMO 81 estimates were calculated for comparison for the NASA/JPL projects. Few of the models met the mean absolute percentage error (MAPE) standard of 25 percent or less; however, managers may find mean error (ME) to be a better metric for evaluating software cost models. ME results for the NASA/JPL projects suggest that managers may prefer the simpler categorical models to the more complex models for smaller programs. The best of these models had 223.4 person-months, or 86.3 percent less error on average than the COCOMO based regression.

Acknowledgments

I want to thank my Lord and Savior, Jesus Christ, for His saving grace and for the ability to complete this educational and life milestone.

I want to thank my loving and patient wife for all her support in accomplishing this goal. I cannot put into words what she means to me or how blessed I am to have her in my life. I thank her for keeping our home running smoothly and everything in check while I was off by myself studying.

I extend special thanks to my thesis advisor, Dr. Michael M. Hicks and my readers, Lt Col Curtis Tenney and Maj Jeffery Smith. I thank them for their guidance that allowed me to complete this project in a manner better than I could have on my own.

Finally, I want to thank the members of my church small group and the other friends and family that supported me with their prayers and best wishes throughout this program.

Marc D. Ellis

Table of Contents

Page

List of Figures

List of Tables

Table | Page

Table Page

A HEDONIC APPROACH TO ESTIMATING SOFTWARE COST USING ORDINARY LEAST SQUARES REGRESSION AND NOMINAL ATTRIBUTE VARIABLES

I. Introduction

Overview

The objective of this chapter is to provide a brief introduction to this thesis by discussing the background, problem statement, research objective, scope, approach and methodology, and the significance of the research. The chapter closes with a preview of the chapters that follow.

Background

"Software spending in the Department of Defense (DOD) and NASA is significant, and it continues to increase (Parametric, 1999:Ch 6, 1)." Increasing software costs directly impact program budgets. It is important that managers and cost estimators have tools that are accurate, easy to understand and explain, and are as inexpensive as possible to execute.

Several key variables used in current software cost estimation models are not available when a new software project is being initiated. For example, size is a primary component of many current software cost models. This construct is often represented by the number of lines of code in a program or the number of functional points associated with the project. While using functional points is an improvement to using lines of code, both parameters can be difficult to estimate at the onset of a new software intensive

project. The best values that can be obtained for these variables are only estimates and have a certain degree of error. Using these estimated variables only adds to the overall error of the prediction models.

Also, to derive these variables, a certain degree of software engineering expertise is required. Often managers and cost analysts depend on subject matter experts such as the software engineers developing the program to obtain the values for these variables. Not only does this dependence on technical experts add to the cost of the estimate, it makes it difficult to obtain a truly independent estimate.

Many of the variables that serve as inputs to current software estimation models provide little meaning to the customer in terms of what he is procuring. This problem is accentuated by the “black box” nature of several proprietary estimation models currently in use. For example, even though variables such as functional points or programming language may be significantly related to software cost, they do not provide the same information to a customer as if variables such as application type or business area were found to be just as significant. Also, even though some of the proprietary cost models may be quite accurate in estimating software costs, explaining and justifying estimates derived from these tools are difficult when the analyst does not understand the input variables and cannot see the calculations that occur within the model.

Problem Statement

The purpose of this thesis is to determine if a hedonic approach using ordinary least squares regression can be used to develop a model that adequately estimates NASA

and JPL software development costs from nominal attribute variables that are more meaningful, cheaper, and or more readily available to the customer.

Scope

This research will evaluate software cost estimation models specifically derived from a set of 60 NASA and Jet Propulsion Laboratory (JPL) software projects. A second set of 116 cases from the International Software Benchmarking and Standards Group (ISBSG) is used to compare and contrast the method and models created for analysis. This research focuses on estimating software costs early in a project's lifetime when little programming and engineering information is available.

Significance

As discussed in the background to this research, software plays a large part in the costs of technology intensive projects. As technology advances, so do the demands for software development and the demands on organizations' budgets. Organizations like NASA and JPL that regularly work with high tech software intensive programs could greatly benefit from any improvement in the software estimation discipline. Any improvement in cost and user-friendly qualities would be valued.

Chapter Preview

This chapter has provided a brief introduction to this research. Chapter 2 documents the literature that was reviewed. The literature illustrates the issues surrounding software cost estimating and supports the methods used to analyze and develop the models derived for this thesis. Chapter 3 outlines the methods used. Chapter

4 provides a detailed explanation of the analysis and findings. Chapter 5 summarizes the research by readdressing the research problem and specific research questions, addresses limitations to the findings, and defines areas of further research.

II. Literature Review

Current Theory and Methods

The text “Estimating Software Costs” explains many of the details of estimating software costs and lays out some of the theory behind several of the current models and methods. Jones says that estimating software costs starts with the size of the project (Jones, 1998:5). Software project size is normally represented by some variant of the number of lines of code that make up the programming, the number of functional points derived from the attributes and capabilities of the project, or a combination of these measures (Jones, 1998:5).

Figure 1 illustrates the fundamental relationships within current software estimation models. This relationship holds for many of the possible software related estimates such as schedule, effort, cost, and deliverables (Jones, 1998:5,6).

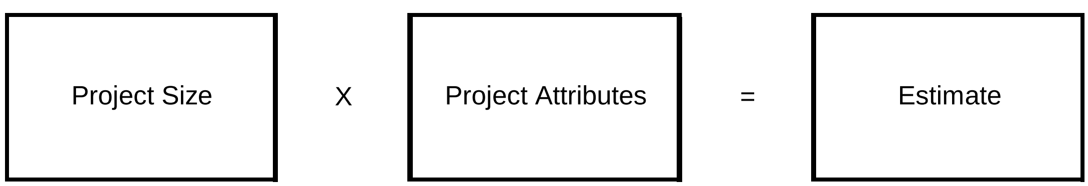

Figure 1. Fundamental Software Cost Model Relationships (Jones, 1998:6)

Effort and cost are used interchangeably throughout the literature. Software development is labor intensive (Lum and Hihn, 2002:1). This labor, or effort, is the primary cost driver for a software project, and many of the models predict effort as the dependent variable (Sommerville, 2004:613,614). Effort is a standard measure that all organizations can use and is not affected by inflation or other factors. Effort is often

measured in person-months or person-years. While effort is a standard measure that can be used across organizations, similar effort measures do not always translate to similar dollars among different companies. This is because each company has different salaries for its programmers and different overhead rates which are also applied over the effort (Sommerville, 2004:613).

Software Cost Estimation Techniques.

The Parametric Estimating Handbook lists and describes four main software estimating techniques: analogy, expert judgment, bottoms-up, and parametric (Parametric, 1999:Ch 6, 19-4). The analogy approach to estimating software costs is comparing a new software project with a similar historical software project that has known cost data (Parametric, 1999:Ch 6, 19). The assumption is that similar projects should have similar cost attributes (Parametric, 1999:Ch 6, 19). While this method is fairly easy and based on historical data, it is often difficult to find a similar project, especially one with historical data (Parametric, 1999:Ch 6, 20).

Expert judgment techniques involve utilizing software program experts' experience, knowledge, and opinion to obtain a best guess estimate of the costs involved in a new project (Parametric, 1999:Ch 6, 20). This method can be helpful when there is no comparable historical data to a new project; however these opinions are subject to bias and not known for being accurate (Parametric, 1999:Ch 6, 20). This method is not used very often for official government estimates (Parametric, 1999:Ch 6, 20).

The bottoms-up technique involves dividing up a new project into as many small parts as possible and through detailed analysis, calculating an estimate for each part (Parametric, 1999:Ch 6, 20). Each individual estimate is then added together to get the

final total estimate (Parametric, 1999:Ch 6, 20). This method can be very time consuming and expensive, and the estimates of each small part are subject to estimating error that is summed into the final estimate (Parametric, 1999:Ch 6, 21).

Parametric methods are the method of choice for government software projects (Parametric, 1999:Ch 6, 21). This method involves statistical relationships between program characteristics and the variables that is being estimated such as cost and/or schedule (Parametric, 1999:Ch 6, 21). Parametric models are usually easy to use and produce estimates at the total system level (Parametric, 1999:Ch 6, 21). The models developed in this thesis employ this method. The project characteristic and program attribute independent variables along with the dependent variable, project effort, are analyzed to determine if any statistical relationships exist. Once can use these statistical relationships with the other methods outlined in the next chapter to develop models to estimate costs based on these input variables. Since the NASA/JPL data considered in this research is from a government source, and parametric methods are the preferred method for government projects, it is appropriate to use this method.

Using Ordinary Least Squares Regression in Software Cost Research.

R. Jeffery, M. Ruhe, and I. Wieczorek conducted a study to examine two questions: Is there a difference in the accuracy of software development cost estimates when using ordinary least squares regression (OLS) on a dataset compared to an analogy-based method, and is there a difference in the accuracy of a cost model derived from data in a public multi-organizational database compared to a model built with company specific data (Jeffery and others, 2000:1009). This paper mentions that OLS is a frequently used technique in software cost estimation research and that analogy methods

have grown in popularity since the 1990s (Jeffery and others, 2000:1009). Jeffery et al. utilized the International Software Benchmarking and Standards Group database, also utilized in the analysis for this thesis, along with a company specific database (Jeffery and others, 2000:1009).

The group concluded that using company specific data produced more accurate models. Also, when using company specific data, both the OLS and analogy method performed well enough to be considered as viable methods (Jeffery and others, 2000:1015). They also determined that OLS performs much better than the analogy method in terms of accuracy when a multi-organizational database is used to create the model (Jeffery and others, 2000:1015).

As described above, size and project attributes are the two primary theoretical parameters when it comes to software cost estimates. In the case of both analogy and parametric methods, project attributes are often used to determine the size of a project (Parametric, 1999:Ch 6, 42,43). While the analogy method depends on direct comparison of historical projects with similar attributes, the methods for determining size for parametric models are often regression based (Parametric, 1999:Ch 6, 42, 43). Considering this close theoretical relationship between project size, effort, and project attributes, and the fact that OLS has been effectively used in prior research, this thesis considers using regression to estimate software effort using only the project attribute variables available in the source datasets. These variables are used not only as predictors of effort, but they serve as proxy variables for project size. Later, this chapter will describe other reasons for excluding the traditional size variables such as lines of code and functional points.

Current Software Cost Models

The Parametric Estimating Handbook lists and describes three commonly used parametric software cost estimation models: the Constructive Cost Model (COCOMO), PRICE Software Model (PRICE S), and the Software Evaluation and Estimation of Resources Software Estimating Model (SEER-SEM) (Parametric, 1999:Ch 6, 21). The next three subsections list and explain these three models in more detail and highlight the important information pertaining to this research.

COCOMO.

COCOMO is a parametric model that estimates software cost utilizing regression and functions in three different nodes designating the development environment: organic, semi-detached, and embedded (Parametric, 1999:Ch 6, 22; Boehm, 1981:78-80). The model predicts the development effort in person-months using the program size as the independent variable (Parametric, 1999:Ch 6, 22). The original COCOMO, also known as COCOMO 81, operates at three levels: basic, intermediate, and detailed. The basic level generates a rough order of magnitude estimate of a software project's cost. The intermediate level includes nominal variables representing characteristics of the software. The detailed level adds to the accuracy of the intermediate level by adding multipliers representing the associated costs of each stage of the development cycle (Parametric, 1999:Ch 6, 22, 23). COCOMO II is an updated version of the model. This model improves upon COCOMO 81 by considering current life cycle practices (Parametric, 1999:Ch 6, 27). Both the COCOMO 81 and COCOMO II models are non-proprietary with all of the internal functions available to the public for review and analysis (Parametric, 1999:Ch 6, 27).

Along with lines of code, the COCOMO 81 model uses input values representing 15 environmental factors. Barry Boehm determined these variables to be significant and independent software development cost drivers (Boehm, 1981:115). They are divided into four groups: product attributes, computer attributes, personnel attributes, and project attributes. The following table lists the four categories and the associated variables in each. These variables and their associated scores were available in the source data used in this research and were analyzed within the NASA/JPL models developed as described in Chapters 3 and 4.

Table 1. COCOMO Environmental Factor Variables (Boehm, 1981:115,116)

Product Attributes	
RELY	Required Software Reliability
DATA	Database Size
CPLX	Product Complexity
Computer Attributes	
TIME	Execution Time Constraint
STOR	Main Storage Constraint
VIRT	Virtual Machine Volatility
TURN	Computer Turnaround Time
Personnel Attributes	
ACAP	Analyst Capability
AEXP	Applications Experience
PCAP	Programmer Capability
VEXP	Virtual Machine Experience
LEXP	Programming Language Experience
Project Attributes	
MODP	Modern Programming Practices
TOOL	Use of Software Tools
SCED	Required Development Schedule

PRICE Software Model.

The Price Software Model, also known as PRICE-S®, is another software cost estimation tool. Unlike COCOMO, some of the equations that make up the model are proprietary and are not available to the public (Parametric, 1999:Ch 6, 33). The PRICE-S® model has nine categories of input information: project magnitude, program application, a productivity factor, design inventory, utilization, customer specifications and reliability requirements, development environment, difficulty ratings for internal and external integration, and development process (Parametric, 1999:Ch 6, 33, 34).

Galorath SEER-SEM®.

SEER-SEM® is also a partially proprietary model, and like the PRICE-S® model, some of the internal equations are not available to the public (Parametric, 1999:Ch 6, 36). SEER-SEM® is used to estimate software cost by inputting values for the software's size, knowledge-base input, and other input parameters (Parametric, 1999:Ch 6, 36). Values for software size come from source lines of code (SLOC), functional points, or some other value representative of the project size. Knowledge-base inputs are preprogrammed values that are activated in the formulas if certain platforms and applications are used. The other input parameters allow the user to refine an estimate by considering the capability of the programmers and the requirements of the software (Parametric, 1999:Ch 6, 36, 37).

The PRICE S and SEER-SEM models are examples of commercial models where several of the internal functions are proprietary and are not available to the public. While these models may be fairly accurate for what they estimate, not knowing all of the

internal workings of these models makes it difficult for the user to completely understand the estimates they create.

Problems with Using Lines of Code and Functional Points as Size Variables

The following three articles describe research that analyzed the effectiveness of the two main variables used to represent size in software cost models: lines of code and functional points. The three articles depict the difficulty of using these variables and the inaccuracy of the models that use them. Due to the errors associated with the use of these variables, one aspect of this research was to design a model that did not use size variables. The aim was to determine if any of the derived categorical variables used in the models would proxy for size in addition to represent other unique effects on effort.

Stamelos et al. stated that there are two major problems with using SLOC as a variable to estimate software development costs. The first problem with this variable is that the true SLOC for a program are not available until project completion, so estimators must predict them during the requirements definition phase (Stamelos and others, 2003:733). The second problem with using SLOC as a variable is that the definition of SLOC is not the same between the different programming styles used by developers (Stamelos and others, 2003:733).

Kemerer tested to see how models utilizing SLOC as an independent variable compared to models that did not use SLOC (Kemerer, 1987:416). The models that did not contain SLOC used functional points or functional-point-like values to represent the size of the project (Kemerer, 1987:418,419). In terms of the magnitude of relative error (MRE), the models without the SLOC variable outperformed the models with it

(Kemerer, 1987:420,427). However, the regression statistics indicated that the models with SLOC as an independent variable outperformed the models without this variable (Kemerer, 1987:427). Kemerer suggests there are problems with this result. One problem is that these models were developed and tested using the actual SLOC values from historical data. The true SLOC will not be available for a new project, so only a prediction can be input into the model for this variable. This decreases the actual predictive power of the model (Kemerer, 1987:427). A second problem is that even though some models that utilize SLOC as a variable make accommodations for the errors associated with predicting this value, it is still unlikely that a technician using the model will have the ability to predict a SLOC value with the accuracy to duplicate the model's predictive power demonstrated in Kemerer's research (Kemerer, 1987:427).

Ian Sommerville also acknowledges that using lines of code or functional points can be problematic to represent size in software cost estimates (Sommerville, 2004:623). Even though functional points frequently outperform lines of code when used to estimate software costs, these estimates are still often inaccurate (Sommerville, 2004:623). Accurate counts for the lines of code are often difficult to determine early in the life of a software project, and the relationship between lines of code and effort are not consistent between programming languages and environments (Sommerville, 2004:616, 623). Sommerville states that there is some subjectivity in regard to the complexity calculations for functional points, and the expert judgment that is required also includes a certain amount of bias and error (Sommerville, 2004:617). Also, functional points are not consistent predictors of size across all system types (Sommerville, 2004:617).

T. Capers Jones states that using functional points requires a certain amount of subjectivity that can lead to complications with the associated estimates (Jones, 1998:351). He adds, "The functional point metrics in all varieties have problems with counting precision for any specific variant, and even larger problems with converting data between variants (Jones, 1998:348)." Variants being functional point calculation techniques. Jones also addresses the problems with using lines of code within software models. He says, "Although LOC metrics were proven to be unreliable for software estimation as long ago as 1978, the technique is still among the most widely utilized (Jones, 1998: 137)."

A Hedonic Analysis of Custom Software

In "Determinants of Price in Custom Software: A Hedonic Analysis of Offshore Development Projects" Ethiraj et al. explain a cost model based on 160 programs developed by a large Indian firm. The study examined estimating software cost for projects outsourced to overseas companies (Ethiraj and others, 2004:2). The research developed a model to predict custom software price by looking at "…specific project, technology, and people related factors that influence the price paid by the client (Ethiraj and others, 2004:4)."

Reviewing prior research, Ethiraj's team found that the price paid for custom software was highly dependent on project characteristics, the conditions and type of contract for the software development, and the agreements made pertaining to quality and service (Ethiraj and others, 2004:6). Project characteristics are the capabilities that the custom software is to be able to accomplish (Ethiraj and others, 2004:6). The team found

that packaged commercial software price is mostly dictated by what customers are willing to buy; however, custom software price is closely related to the development cost. Thus characteristics that are the major cost drivers in custom software development must be considered in the model (Ethiraj and others, 2004:7).

Contracts can affect the price of a custom software project by the way they determine who is responsible for costs of development when estimates are exceeded (Ethiraj and others, 2004:8). If the purchaser is responsible for overages, they will get a product, but there is a chance that the vendor may be more likely to be less responsible with resources knowing they will not have to pay the extra cost. If the contractor is responsible for overages, there is the chance that the purchaser will either not get a final product or will receive a product different from expected (Ethiraj and others, 2004:8).

The final variable they considered for its effects on cost, and indirectly price, is the quality of software. The researchers assert that it is common knowledge that the level of quality in a project or operation will drive the cost. Their review of research in the area of custom software provided no evidence to the contrary for this type of project (Ethiraj and others, 2004:9, 10).

The research team used ordinary least squares regression utilizing the whitewashing technique to develop their pricing model (Ethiraj and others, 2004:23). The dependent variable, price, was represented by the revenue generated from the project. The natural logarithm of the revenue was the actual variable used in the model. This decision was made due to the lack of theory available for this type of research, and similar methods were used in earlier studies (Ethiraj and others, 2004:23).

The independent variables represented the project characteristics, the mode of the contract, and the quality measures required of the finished project. The project characteristics were defined by nine variables: size and complexity, development team size, project duration, contractor experience, project team experience, education dispersion of developers, industry, program platform, and time (Ethiraj and others, 2004:18-21).

Ethiraj et al. represented contract type by one binary variable that indicated whether the contract was time and materials or fixed price (Ethiraj and others, 2004:21). The variables in-process defect density, effort overrun, and schedule slippage represented project quality (Ethiraj and others, 2004:22).

The team determined the statistical significance and magnitude of each of the tested variables in relation to the dependent variable. The direction of the relationship was also discovered. However, the team did not provide much information on the predictive quality of the model developed (Ethiraj and others, 2004:24-28).

NASA and JPL software projects are usually custom software programs, and like the programs discussed in the research of Ethiraj et al., the price of the software is closely related to the cost of development. The hedonic approach to building and analyzing a model that predicts price for these Indian developed custom programs could be applied to a cost model due to this similarity. The analysis in this thesis attempts to use a similar approach to determine project characteristics that are good predictors of project costs and that provide meaning to the user due to the descriptive value of the characteristics and their statistical relationships.

Project Attributes in Software Cost Models.

Stamelos et al. consider estimating development costs of custom information system software in their article "Estimating the Development Cost of Custom Software". The group says that while purchasing commercial off-the-shelf programs is increasing, there will always be a need for estimating the costs of custom software for the specific needs of some companies (Stamelos and others, 2003:729,730). In Software Cost Estimation (SCE), one of the most important cost drivers are the costs associated with the human effort involved in the software development (Stamelos and others, 2003:730). In this research, the team equates software development cost with human effort and develops a model to predict cost using a combination of analogy, functional points, and the calibration of the model to a specific dataset (Stamelos and others, 2003:730, 739).

The group combined the results of these three measures to produce the final cost estimate. The best estimate produced using this method had a mean magnitude relative error (MMRE) of 23.84% (Stamelos and others, 2003:738). "An acceptable target value for MMRE is 25%, meaning that an average project may be estimated with a relative error of 25% with respect to its actual effort (Stamelos and others, 2003:730)." Combining these three methods to estimate software development cost appears to produce acceptable estimates, and since SLOC are not used as a variable, the method can be used as early as the requirements definition phase (Stamelos and others, 2003:739).

Stamelos et al. described research that clearly laid out several of the main theoretical aspects of software research and how they could be applied to developing and analyzing a new predictive model or method. They discussed the basics of the analogy technique, sizing methods, and the relationship between size and effort (Stamelos and

others, 2003:732-734). They also provided evidence that comparing project attributes in a model could produce acceptable software effort estimates. While not necessarily an analogy-based model like the one produced by Stamelos et al., the variables in the models produced for this thesis are calibrated against similar attributes from other projects. However, the models in this research are regression-based rather than analogy-based. Stemelos et al. also provided another example of research that utilized the MMRE (also known as MAPE) standard of 25 percent or less as a standard for acceptable models. They also used the ISBSG dataset.

L. Angelis et al. studied the possibility of building a cost model using a categorical regression technique. This technique used variables representing qualitative characteristics of the software to be developed to predict the cost of the project (Angelis and others, 2001:4). The research team used the ISBSG database for their analysis. The ISBSG database is a collection of data from hundreds of software projects and is available for research, software benchmarking, and cost estimation (Angelis and others, 2001:4-6). Many of the variables in the database are categorical in nature, which led the team to test the possibility of using these variables to predict the software cost of a new project by its categorical attributes (Angelis and others, 2001:4, 9).

The dependent variable in the model was the natural log of effort (Angelis and others, 2001:5, 7). The groups first developed a linear regression model with the natural log of the effort (LN(Effort)) predicted by the natural log of the number of functional points provided in the data set for each program (Angelis and others, 2001:7, 8). This model showed only a weak relationship (R^2 of .44) (Angelis and others, 2001:9). They then used their categorical technique by adding categorical variables to the model to

improve its prediction value. The categorical values were chosen according to previous research indicating variables with strong relationships with development effort. They tested each variable using the one-way ANOVA test, and statistically significant variables were included in the model (Angelis and others, 2001:9).

Eventually, seven categories were chosen to add to the model, and combinations of particular variables representing these categories were combined into the model and were analyzed. The categories were development type, development platform, language type, use of development methodology, organization type, business area type, and application type (Angelis and others, 2001:8).

A nominal variable indicating either an enhancement project or new development represented the development type category (Angelis and others, 2001:10). A nominal variable indicating either development on a personal computer or on mid-range to mainframe systems represented the development platform category (Angelis and others, 2001:10).

Language type was a nominal variable indicating either a low language (4th generation) or a high language (all others) (Angelis and others, 2001:10). A nominal variable indicated whether the development of a particular software program used an innovative methodology (Angelis and others, 2001:10). The authors of this study did not provide much insight into how an innovative methodology was defined. This variable was included in the ISBSG database; however, the database did not provide the researchers with a good idea of how this variable was determined (Angelis and others, 2001:10).

The organization category was represented by a nominal variable indicating one of two groups of industries. The first group was services and administration. The second group was finance, manufacturing, and operations (Angelis and others, 2001:10). Business area type was defined by a nominal variable indicating one of two levels. Angelis et al. considered research and development, telecommunications, engineering, sales, and financial programs as low-level business types, and banking, accounting, legal, personnel, manufacturing, and inventory as high-level business types (Angelis and others, 2001:10).

Finally, application type was a nominal variable indicating two options. One was for high-level applications such as transaction/productions systems and office automation systems. Process control, network management, management information systems, executive information systems, electronic data interchange and decision support systems were then indicated to be low-level applications (Angelis and others, 2001:10).

Angelis et al. initially considered the individual parts of several of these categories as separate inputs, but they combined them into groups that showed significant relationships to the dependent variable (Angelis and others, 2001:9, 10). Angelis et al. suggest that the categorical model performs better at predicting software costs from the data compared to a simple linear regression of the same data set (Angelis and others, 2001:14). However, they were unsure what variables were included for this comparison. They also do not indicate the degrees of superiority nor does it state how it was determined. Angelis et al. also stated that the results of the categorical model could not be tested against other research utilizing categorical variables since the actual variables

available in different databases would be quite different making a one-on-one comparison of categorical models very difficult (Angelis and others, 2001:14).

Angelis et al. provide support for considering categorical attribute variables in software cost models. A different categorical approach was used on the ISBSG database and a database of NASA and JPL projects for this thesis. Different variables, variable combinations, and selection techniques were used. However, some of the same categories and attributes from the ISBSG dataset were considered in the models for this thesis. Also, the natural log transformation of the dependent variable was also tested in this thesis. The technique used by Angelis and his team for variable selection did not allow any statistically insignificant variables to enter the model. Valuable information can be lost when variables are eliminated based on their significance in a model. Once the variable set for analysis had been selected for this thesis, the variables were only eliminated when singularity and multicollinearity issues were present. While there were several differences between the research done by the authors of this paper and the research accomplished for this thesis, the idea of using categorical variables to estimate software cost is the same.

Other Research Using the NASA/JPL Data.

In "The Structure of the Software Cost Function," Karen T. Lum and Jairus Hihn state that current software cost models are not modeled after the commonly accepted cost and production functions (Lum and Hihn, 2002:2). Their aim was to test whether or not wage rate, an input into the standard production function, is statistically significant when it comes to estimating software cost (Lum and Hihn, 2002:2). While theory states that

wage rate should be considered, their analysis found that this variable was statistically insignificant (Lum and Hihn, 2002:9). There was insufficient evidence to reject the current cost estimating relationships built in the existing cost models (Lum and Hihn, 2002:8). Their article is one of the two that document research using the source dataset used in this thesis. Their document was the source for most of the NASA/JPL data descriptions described in Chapter 3.

In "Economic Analysis of the Software Cost Function," Lum and Hinh describe research that is an extension of the study described above. In their second study, they are again testing to see if the standard production function, especially wage rate, is significant in software development effort (Lum and Hihn, 2004:1,2). This time they added over 60 cases to the dataset used in the previous study and normalized their data for inflation (Lum and Hihn, 2004:4). They discovered that wage rate was significant for flight software; however, this variable remained insignificant when considering ground software projects (Lum and Hihn, 2004:7). The only variables significantly related to the ground projects were size variables and the tested environmental variables (Lum and Hihn, 2004:7).

Using Mean Absolute Percentage Error as a Model Evaluation Metric.

In the article "A Simulation Study of the Model Evaluation Criterion MMRE," Foss et al. discuss their research that evaluates using mean magnitude of relative error (MMRE) as the standard error measure for evaluation software cost models (Foss and others, 2003:1). MMRE is the most recognized and accepted metric used to evaluate and compare software cost models (Foss and others, 2003:1). The accepted standard for this

measure for an adequate model is 25 percent and below (Foss and others, 2003:1). Foss and his colleagues say that while MMRE is very common in the software and computer science disciplines, it is less known and utilized in other disciplines. However, they do acknowledge its use in evaluating time series forecasts. They reference Makridakis et al. who refer to this measure as the mean absolute percentage error (MAPE) (Makridakis and others, 1998:44,45). Mean percentage error (MPE) is the average of all percentage errors calculated for each prediction. One calculates MAPE the same as MPE except that it is an average of the absolute values or magnitudes of the percentage errors.

Foss and the other researchers simulated 1,000 software projects and the parameters necessary to develop and use different cost models to calculate estimates for the effort. The team tested how well each model estimated the effort in comparison to the known effort values using MMRE, and other evaluation metrics. They found that MMRE does not consistency choose the best model to estimate effort. They found that MMRE actually favors models that underestimate effort.

MAPE was the primary evaluation metric used when comparing the models generated in this thesis. Despite the weaknesses illustrated in the study by Foss et al., MAPE is one of the few error measures that provides a meaningful score representing the relative magnitude of the estimate errors for easy comparison (Makridakis and others, 1998:43-45). While other metrics may select the appropriate model more frequently, it is difficult to determine from these measures how well one particular model performs on a set of data with their measures. MMRE provides this value since prior research supports the standard that scores for this measure equal to 25 percent or less are acceptable. Foss et al. demonstrate the problems with using MAPE, yet they admit that they do not have

one single measure that works best in all cases (Foss and other, 2003:993). They also did not provide information on what scores would be acceptable standards for the different metrics tested. MAPE remains the standard measure to evaluate and compare software cost estimates in most of the research literature, and the analysis in this thesis uses it as well.

Foss et al. listed other items of interest to this research. They described how software data for cost estimation often displays heteroskedasticity in their values (Foss and others, 2003:987). It also describes how natural log transformations are used to attempt to reduce the effects of heteroskedasticity in order to meet OLS assumptions (Foss and others, 2003:989). Testing the residuals of several of the models in this thesis revealed non-constant variance. Like Foss et al., the natural log of the dependent variable, effort, was tested for each model type described in Chapter 4.

A Final Note on Software Cost Model Accuracy

Ian Sommerville brings up an interesting point in regard to software cost estimates and their effect on actual costs. He says, "Project cost estimates are often self fulfilling. The estimate is used to define the project budget, and the product is adjusted so that the budget figure is realised [sic] (Sommerville, 2004:620)." This makes testing new methods and models on historical data very difficult. True relationships between variables may not be identified due to the bias involved. Current models may perform at the level they do only because contracts were completed based on this effort estimate rather than of any other model estimate or on actual effort to complete the project.

III. Methodology

Data

The analysis in this thesis considered two different databases. The first database is a small compilation of National Aeronautics and Space Administration (NASA) and Jet Propulsion Laboratory (JPL) software development projects. The second set of data is a commercial database provided by the International Software Benchmarking and Standards Group (ISBSG).

NASA/JPL Data.

The NASA/JPL dataset is composed of 60 ground software projects that were developed between 1986 and 1990 (Lum and Hihn, 2002:4). The database lists lines of code, actual effort, average monthly wage rate, a derived actual cost based on the effort and wage rate, environmental factor-level ratings for 15 COCOMO variables, and a labor multiplier based on these environmental factors (EAF).

The analysis of this thesis did not use the monthly wage rate and derived cost data. Actual effort was the dependent variable in this study representing cost rather than the derived dollar value in the source data. Effort is commonly used in software cost analysis and estimation as demonstrated in the literature review. Lines of code were recorded in thousands. Effort was recorded in person-months. A person-month is one month excluding holidays, vacations, and weekends, worked by one person. Total person-months for each project is the sum of all months worked by all individuals on a software development project. (Lum and Hihn, 2002:4,5) The 15 environmental factors are the input variables for the COCOMO 81 software cost model. Their corresponding

multipliers as required in the model represented these variables. Lum and Hihn interviewed subject matter experts involved in the projects listed in the dataset and asked them to rate each of the COCOMO variables on the appropriate scale for each of the 60 projects (Lum and Hihn, 2002:4).

The EAF labor multiplier relates the total effect of the COCOMO factors to labor. Multiplying each of the numerical values for the COCOMO input variables together equals the EAF (Lum and Hihn, 2002:4).

The data source also included short descriptions of the overall program, the parent project, the program element, and the software category for each case. These descriptions were used to derive six general categorical variables that would represent these factors. The six variables are data capture/collection, data processing, avionics monitoring, simulation, mission design/sequencing/planning, and command and control.

Binary dummy variables represent the six variables derived from the project descriptions. For each of the 60 projects, the number 1 represented each category or attribute description that pertained to that particular case. Zeros for each of the variables represented projects that did not match any of these six variables.

The source dataset contained the appropriate multiplier values for each of the variables necessary to run the COCOMO model. However, while these values are numerical, they are nominal in nature because each value represents a specific rating on a scale between very low to extra high for each input. For this thesis, each environmental factor and its appropriate multiplier levels were split into separate variables. Then for the entire set of 60 cases, a “1” or a “0” represented each factor level if it pertained or not.

Preliminary analysis indicated a great contrast in effort between programs under and above 500 person-months. This threshold value was used to separate projects into two size categories for analysis and comparison. For each of the 60 cases, a size variable was recorded. A "1" was assigned to all cases with over 500 person-months of effort, and a "0" was assigned to all cases with under 500 person-months of effort. These values enabled separating the cases by size for different models.

ISBSG Data.

The second data source used was a collection of software development data maintained by the International Software Benchmarking Standards Group (ISBSG). The Repository Data CD, Release 8 version of this database contains over 2,000 software projects from over 20 countries with over 75 percent of them being developed within five years of its release in 2003 (What, 2003:1). Several software cost and development studies use the ISBSG dataset.

Within the ISBSG dataset, each project has an assigned quality rating "A", "B", "C", or "D". An "A" rating represents data that has no identified quality or integrity issues. Ratings from "B" through "D" indicate data with increasing data quality problems (Repository, 2003:2). This thesis only considers projects with "A" ratings.

Effort in this database is recorded in person-hours (Repository, 2003:1). The type of program tasks included in the count for total project effort is not consistent throughout the database. Each project counts person-hours from some combination of project planning, specification, design, building, testing, and implementation (Repository, 2003:4). For the sake of consistency, this research only considered projects when all six phases were recorded. By limiting this research to projects with an "A" quality rating

and to those with effort counted as previously mentioned, this thesis analyzed 116 projects within the final working database from this source.

Eight categories were chosen from the ISBSG dataset that could be clearly separated into variables and recorded with dummy values. The categories included were as follows: application type, business area, development platform, development technique, development type, programming language, language type, and organization type. Each of these categories had several types or subcategories. For example, three of the application types within the application type category were graphical user interface (GUI) application, management information system, and decision support system. The numbers 1 and 0 represented whether a particular attribute was present for each category for model development and analysis.

Actual implementation dates for most of the projects were also available in this database. This variable was recorded by day, month, and year. This information was modified for analysis by converting the date for each case into the number of years since the date of the earliest implementation date within the 116 projects. Functional point values for each case were also available. Like the NASA/JPL data, preliminary analysis suggested a variable should be added to separate the cases into two levels of effort. The numbers 1 and 0 represented projects over and under 15,000 person-hours respectively. This variable allowed for separation of the cases for analysis by the different sizes.

Variable Selection and Model Development

The NASA/JPL models are the primary focus of this research. The models derived using the ISBSG data were calculated to compare how the methods and models

perform on a different, less specific set of data. The primary tool for developing the models was ordinary least squares regression.

In the case of both datasets, there were many possible variables that could be considered in the models. Stepwise regression was initially considered as a method for selecting variables and developing models from the available data. However, stepwise regression is not always a well regarded method. Peter Kennedy describes how this method maximizes the R^2 value of a regression by adding and subtracting variables in the model without regard to theory or the nature of the environment modeled (Kennedy, 2003:58). A more structured and systematic approach was applied in this research. Correlations with the project effort guided initial consideration of variables within each category. The next chapter provides a more detailed explanation of how the variables were selected. This method allowed for each type of model tested for each dataset to be analyzed using the same variables.

Model Analysis

Each of the models was reviewed according to its level of fit and statistical significance. In additions, each model was tested in regard to meeting the assumptions for OLS regression. After a model was developed using the random sample of the available cases, the model was tested by estimating the effort for the reserved projects. Five error measures represented the predictive qualities of the developed models when tested on the reserved data: mean error (ME), mean absolute error (MAE), mean percentage error (MPE), mean absolute percentage error (MAPE), and mean squared

error (MSE). Particular attention was given to the MAPE score for each model as a score of 25 percent or less is a widely accepted standard for a suitable software cost model.

The initial model of each type did not include lines of code and functional points due to the problematic issues with using these variables as discussed in the literature review. However, the data for these variables were available in the source datasets. Lines of code were added to the NASA/JPL models and functional points to the ISBSG models to compare how these models would perform if this information was included. The performance of these adapted models would be subject to the availability and accuracy of these variables for future project estimates.

The variables were available in the NASA/JPL source dataset to actually calculate estimates using the COCOMO 81 model. These estimates were calculated and recorded to provide another means of analyzing and comparing the results produced by the models developed. The actual COCOMO model requires lines of code in order to calculate effort. Comparing the predictive qualities of the models developed in this thesis that do not incorporate lines of code to the COCOMO estimates is not a satisfactory comparison. Two regressions, one with a semi-log transformation of effort and the other without, were developed using the actual input values for each of the COCOMO environmental factors but no lines-of-code value for a truer comparison.

IV. Analysis and Findings

Model Development and Variable Selection

The analysis in this thesis tests two different sets of models using two source datasets. The first set of models used the NASA/JPL data, and the second set used the ISBSG data.

NASA/JPL Model Development.

After breaking down each COCOMO factor into separate binary variables representing the different factor-levels, 56 different variables were available for consideration in the model. Since Boehm's previous research indicates that all 15 COCOMO factors are significant and independent cost drivers, it is reasonable that each factor be represented in this research (Boehm, 1981:115). However, including all 56 variables is inappropriate, if not impossible, for models developed from this small set of 60 cases. It was necessary to determine a way to choose which variables would be included in the model. Stepwise regression was initially considered, but this method was not chosen for reasons stated in the previous chapter. Instead, a more relevant method that considers the nature of the variables and the data was devised. Correlations were calculated for each COCOMO factor-level in regard to effort. The factor-level with the greatest correlation in terms of magnitude for each COCOMO category was chosen as the variable to represent that category. For example the low complexity factor level had the greatest correlation of all of the complexity factor levels, so low complexity (CPLXL) was chosen as the variable to represent this COCOMO factor. Fifteen factor-level

variables were chosen this way. Appendix A lists the factor-level variables chosen to be included in the models utilizing this dataset.

All six of the project descriptive variables were considered in the models. The only time any of these six variables or the 15 COCOMO factor-level variables were removed from a model was to correct multicollinearity and singularity issues when identified.

A sample of 48 cases, 80 percent of the database, was selected for model development. Each of the NASA/JPL models analyzed later in this chapter was developed from this random sample. The models that were created were then tested on the remaining 12 cases or 20 percent of the database. Ordinary least squares regression was the chosen tool for model development. Several combinations of variables based on type and size were built into models for analysis.

ISBSG Model Development.

Since stepwise regression was eliminated as a method for model development, a structured and systematic approach similar to the one used for the COCOMO factor-levels was used to develop models from the ISBSG data. Correlation matrices were completed for each category. Binomial values represented the presence, or lack thereof, of each possible characteristic in each category. Using the correlation matrix, correlations between the independent variables were ranked by size. Starting with the highest correlation between variables, the variable with the lowest correlation to effort was removed. The correlations were then re-ranked and the process was repeated until no correlations between variables were greater than the correlation between either variable involved and the effort. For example, within the application type category, only

transaction/production systems and graphic user interface applications had correlations to project effort greater than correlations to other potential independent variables in this category. These were the two variables from this category included in the models for analysis.

Since the overall objective of this research was to determine if an effective cost model could be developed from categorical variables that were easily and cheaply available at the initial stages of a software project, categories were selected from the ISBSG source data that were assumed to meet this criteria.

The modified time component was regressed against the effort to identify the magnitude and significance of this variable. This variable was statistically significant, yet it represented less than four percent of the variance in the data. Since there was little relationship between this variable and effort, and time was not a categorical type variable, which was a main consideration in this research, this variable was not included in the models for analysis.

A random sample of 80 percent of the 116 cases was selected for model development. This amounted to 93 cases. The remaining 24 cases were used to test the predictive qualities of the models. OLS, sometimes with a semi-log transformation of the dependent variable, was the method used to create the models. Several combinations of variable types and sizes were analyzed.

Singularity and Multicollinearity

After all of the models were created, some singularity and multicollinearity issues had to be resolved. The statistical software indicated variables with singularity issues by

labeling them as “zeroed” in the output of parameter estimates. These variables were then removed, and the regression was recalculated. Initial multicollinearity issues were identified using variance inflation factors (VIF). A variable with a VIF over 10 indicates that multicollinearity issues are present that could be damaging to OLS outputs (Kennedy, 2003:213). Of all variables with VIF scores over 10, the variable with the largest p-value was removed. The regression was recalculated, and VIF scores were checked. The process was repeated until all VIF scores of the remaining variables were less than 10. This process was not failsafe as further multicollinearity issues were present in some models when a particular model was overall statistically significant, yet few or none of the parameters were statistically significant (McClave and others, 2005:882). However, there is opinion that if multicollinearity exists in the environment the model is representing, it is inevitable that it will remain in the model. One view handling this issue is to leave collinear independent variables in the model (Kennedy, 2003:210). This was the approach taken when multicollinearity was still evident even after the variables with VIF scores over 10 had been removed.

Testing OLS Assumptions

There are four primary assumptions that the residuals from a multivariate regression model should meet: a mean of zero, constant variance for all independent variables, a normal distribution, and independence (McClave and others, 2005:891). To test for a mean of zero, a 95 percent confidence interval for the mean was calculated. If this interval included zero, there is insufficient evidence to indicate that the error mean is

not zero. The Breusch-Pagan test was used to test the constant variance assumption. The Shapiro-Wilk W test was used to test the distributions of the residuals.

Care was taken to insure that for each dummy variable only the presence, and not both the presence and the absence for a particular attribute, was represented by a dummy variable. In instances where several attribute possibilities were available for a particular category, at least one attribute or the unknown was represented by the absence of all other attributes (i.e. zeros were recorded for all associated variables). If this is not done, a perfect relationship exists between the presence of some attributes and the absence of others, thus violating the assumption of independence. Also, other independence violations were identified and resolved when singularity issues were identified between two or more of the independent variables.

Analysis of NASA/JPL Models

Three separate tables are provided for each of the three groups of models in the following sections. The models in the first group are those derived from all of the cases in the sample. The models in the second set are derived from those cases with less than 500 person-months of effort. The models in the third set were derived from only those cases with over 500 person-months of effort. The first table in each set shows what combination of project descriptive variables and COCOMO factor-level and EAF variables were included. The first table also displays five error measures, an R^2 value, a p-value for each model, and whether or not a natural log transformation of the dependent variable was employed. It also indicates if the OLS assumptions were met for each model.

The second table in each group displays the top five variables according to their p-values for each model. If a variable among the five is statistically significant (p-value < 0.05), it is represented by a number in bold and italicized font. An "X" in the LOC line indicates that the lines of code were statistically significant as an input variable when included, and an "O" indicates that this variable was not significant. The letter "N" indicates that lines of code were tested for a particular model, but the results were erroneous due to an infeasible perfect representation of the effort by the model. The third table for each group displays the top five variables according to magnitude within each of the models. Again, if one of these variables happens to be statistically significant, it is indicated with bold and italicized font.

For both the second and third tables, the range of values for each model is between 1 and 5. The number 1 indicates either the smallest p-value or greatest magnitude, and a 5 represents the 5th smallest p-value or 5th largest magnitude. The letter "S" and "L" under each model number indicates whether it is the standard model or if lines of code are included respectively. The dashed line separates the six project descriptive variables from the COCOMO factor-level and EAF variables.

Models one through six are developed using 48 randomly selected cases from the database of 60 cases. Models seven through 12 are developed using only those cases with 500 person-months of effort or less from the 48 randomly selected cases. Models 13 through 18 are developed using only those cases with over 500 person-months of effort from the randomly selected 48 cases. Lines of code are added as a variable to each model for purposes of comparison. Each model is tested on the like variables of the reserved 12 cases left from the total sixty.

Model 4 and Model 6 were the only two models based on the NASA/JPL data that were statistically significant and met the OLS assumptions. Their model numbers are in bold italicized font in the second and third tables.

The models and regression statistics derived from the cases with over 500 person-months are very limited in regard to statistical standards. There were only 11 cases over 500 person-months of effort within the 60-case dataset. Only 9 were randomly selected for model development, and two remained in the reserved set for testing. While the results from these tests are limited, they may provide some insight into any differences between the large and small projects.

The two functions below represent the basic form of the NASA/JPL models analyzed. Each model contains some combination of the six project descriptive variables and the COCOMO factor-level and EAF variables. Lines of code were added to each model for testing, and a natural log transformation of the dependent variable was tested for each independent variable combination.

Effort in Person-months = f(six project descriptive variables, 15 COCOMO based factor-level variables and EAF, lines of code) (1)

Ln (Effort in Person-months) = f(six project descriptive variables, 15 COCOMO based factor-level variables and EAF, lines of code) (2)

NASA/JPL Model Group 1.

According to Table 2, it appears that most of the models in this group are statistically significant with strong R^2 values. However, all of the MAPE scores fail to meet the accepted standard of 25 percent or less. Not only do the models fail to meet this standard, the magnitude of the failure for most of them is very large. The ME values

indicate there was a mixture of over and under predicting effort by the models on average. The models with natural log transformation of the dependent variable outperformed models without this transformation. There does not appear to be any relationship between meeting the OLS assumptions and better MAPE scores, R^2 values, or p-values. Two of the worst MAPE scores occur when the six project characteristic variables are included in the models. One model included the COCOMO factor levels and EAF, and the other did not. Including lines of code greatly improved each model in regard to its MAPE scores.

Table 2. Performance Summary for NASA/JPL Models, Group 1

All Cases (All Sizes)											
	Project Characteristic Variables	COCOMO Factor-Level Variables	Ln (Effort)	ME	MAE	MPE	MAPE	MSE	R^2	P-value	Meets OLS Assumptions
Model # 1	X	X		105.20	438.60	-430.10	526.50	612,840.10	0.73	0.0003	No
With LOC	X	X		-167.10	261.20	-77.20	149.00	153,323.30	0.98	< 0.0001	No
Model # 2	X	X	X	255.15	323.74	-64.84	124.00	516,306.53	0.62	0.017	No
With LOC	X	X	X	-56.12	124.17	-22.66	42.73	53,949.34	0.84	< 0.0001	No
Model # 3	X			-51.83	512.65	-650.03	675.99	565,643.63	0.21	0.129	No
With LOC	X			-29.11	216.92	-62.01	159.93	121,927.60	0.90	< 0.0001	No
Model # 4	X		X	207.42	362.68	-162.99	218.82	488,749.93	0.29	0.024	Yes
With LOC	X		X	102.63	151.23	-48.27	78.06	91,386.56	0.78	< 0.0001	Yes
Model # 5		X		53.96	391.60	-203.33	255.04	437,823.34	0.68	0.0004	No
With LOC		X		-115.37	184.03	-4.50	118.68	101,695.45	0.97	< 0.0001	No
Model # 6		X	X	130.73	383.25	-79.16	115.15	566,246.48	0.58	0.0093	Yes
With LOC		X	X	-28.56	240.92	-61.28	72.80	251,976.92	0.81	< 0.0001	Yes

Table 3 shows that when only the project descriptive variables are included, simulation has lowest p-value in most cases. When both these six variables and the COCOMO factor levels are included in the models, simulation only makes the top five in one model. The low reliability and low complexity factor levels appear to be significant in several instances whether the six project descriptive variables are included or not.

Table 3. Variable Significance, NASA/JPL Models, Group 1

All Cases (All Sizes) - Variable Significance												
NASA/JPL Model Number	1	1	2	2	3	3	**4**	4	5	5	**6**	6
Standard (S), with Lines of Code (L)	S	L	S	L	S	L	**S**	L	S	L	**S**	L
MissDSP	***1***	--	--	2	4	***2***	--	--	--	--	--	--
ComContrl	--	--	2	--	2	5	4	4	--	--	--	--
DataCC	--	--	5	--	3	4	2	***2***	--	--	--	--
DataP	--	--	--	3	5	--	3	3	--	--	--	--
SimI	--	--	--	5	***1***	***1***	***1***	***1***	--	--	--	--
AvMon	--	--	--	--	--	3	5	5	--	--	--	--
RELYL	***2***	--	--	***1***	--	--	--	--	***2***	--	--	***1***
CPLXL	4	***1***	***1***	4	--	--	--	--	--	***1***	3	3
TURNH	3	***3***	--	--	--	--	--	--	5	***4***	--	--
ACAPL	5	--	--	--	--	--	--	--	--	--	--	--
VIRTH	--	***2***	--	--	--	--	--	--	--	***5***	--	5
LEXPN	--	***4***	--	--	--	--	--	--	--	***3***	5	4
PCAPVL	--	***5***	4	--	--	--	--	--	--	***2***	4	--
STORH	--	--	3	--	--	--	--	--	***1***	--	1	--
EAF	--	--	--	--	--	--	--	--	3	--	--	--
VEXPL	--	--	--	--	--	--	--	--	4	--	--	--
DATAL	--	--	--	--	--	--	--	--	--	--	2	2
TOOLN	--	--	--	--	--	--	--	--	--	--	--	--
TIMEH	--	--	--	--	--	--	--	--	--	--	--	--
MODPH	--	--	--	--	--	--	--	--	--	--	--	--
SCEDVH	--	--	--	--	--	--	--	--	--	--	--	--
LOC	--	X	--	X	--	X	--	X	--	X	--	X

Table 4 shows that the simulation variable also has the largest magnitude when only the six project descriptive variables are included. The low reliability and low complexity factor levels also appear to have large magnitudes in relation to the other variables when the project descriptive variables are included in the models and when they are not.

Table 4. Variable Magnitude, NASA/JPL Models, Group 1

All Cases (All Sizes) - Variable Magnitude												
NASA/JPL Model Number	1	1	2	2	3	3	**4**	4	5	5	**6**	6
Standard (S), with Lines of Code (L)	S	L	S	L	S	L	**S**	L	S	L	**S**	L
MissDSP	***3***	--	--	4	4	***2***	--	--	--	--	--	--
ComContrl	--	--	4	--	2	5	3	4	--	--	--	--
DataP	--	--	--	5	5	--	4	3	--	--	--	--
Siml	--	--	--	--	***1***	***1***	***1***	***1***	--	--	--	--
DataCC	--	--	--	--	3	3	2	***2***	--	--	--	--
AvMon	--	--	--	--	--	4	5	5	--	--	--	--
RELYL	***1***	--	5	***1***	--	--	--	--	***1***	3	--	***1***
CPLXL	2	***1***	***1***	2	--	--	--	--	--	***1***	1	2
STORH	4	--	2	--	--	--	--	--	***2***	--	2	--
TURNH	5	--	--	--	--	--	--	--	--	--	--	--
VIRTH	--	***2***	--	--	--	--	--	--	4	***2***	--	4
TURNH	--	***3***	--	--	--	--	--	--	--	--	--	--
PCAPVL	--	***4***	3	--	--	--	--	--	--	***4***	3	--
LEXPN	--	***5***	--	--	--	--	--	--	--	***5***	--	5
VEXPL	--	--	--	3	--	--	--	--	3	--	--	--
EAF	--	--	--	--	--	--	--	--	5	--	--	--
VEXPL	--	--	--	--	--	--	--	--	--	--	4	--
DATAL	--	--	--	--	--	--	--	--	--	--	5	3
TOOLN	--	--	--	--	--	--	--	--	--	--	--	--
TIMEH	--	--	--	--	--	--	--	--	--	--	--	--
MODPH	--	--	--	--	--	--	--	--	--	--	--	--
ACAPL	--	--	--	--	--	--	--	--	--	--	--	--

NASA/JPL Model Group 2.

Several models in Table 5 also appear to have significant p-values with strong R^2 measures. However, all these models fail to meet the standard for MAPE. Many of these models over-predict effort on average. The models with the semi-log transformation of effort outperform models without this transformation. Only two of the models in this group meet the OLS assumptions, but these models do not appear to perform any better or worse than the models that do not meet these assumptions in regard to MAPE, R^2, or p-value. Adding the lines of code improved all but one of the models in regard to MAPE values.

Table 5. Performance Summary for NASA/JPL Models, Group 2

Cases < 500 Man-Months of Effort												
	Project Characteristic Variables	COCOMO Factor-Level Variables	Ln (Effort)	ME	MAE	MPE	MAPE	MSE	R^2	P-value	Meets OLS Assumptions
Model # 7	X	X		-31.78	99.79	-140.85	165.35	12,890.96	0.35	0.4669	No
With LOC	X	X		-42.64	55.31	-30.50	61.84	6,302.53	0.92	< 0.0001	No
Model # 8	X	X	X	18.27	79.51	-58.29	93.22	11,746.37	0.32	0.559	No
With LOC	X	X	X	-185.83	222.90	-59.70	84.13	332,143.78	0.80	< 0.0001	No
Model # 9	X			-35.41	125.97	-200.09	228.25	20,126.42	0.15	0.478	No
With LOC	X			-35.47	64.80	-26.55	65.59	8,280.10	0.88	< 0.0001	No
Model # 10	X		X	11.37	110.97	-109.66	149.47	19,932.63	0.22	0.2	Yes
With LOC	X		X	-181.60	214.52	-69.46	83.83	322,189.42	0.76	< 0.0001	No
Model # 11		X		-88.36	114.38	-191.04	203.14	16,408.68	0.88	0.57	Yes
With LOC		X		-61.69	69.02	-63.78	66.69	10,622.82	0.90	< 0.0001	No
Model # 12		X	X	-77.41	106.92	-85.39	99.12	24,828.50	0.24	0.67	No
With LOC		X	X	-357.22	394.88	-101.67	120.98	1,231,686.98	0.79	< 0.0001	No

In Table 6 it appears that the command and control project descriptive variable has one of the lower p-values when all variables are considered in this group of model. When only the project descriptive variables are considered, the simulation variable has one of the lowest p-values. The COCOMO factor level variables do not have consistent p-value ratings across the models in which they are included. When only the COCOMO factor-level variables are used, the low-database-development-size variable is in the top five variables according to p-value in each of the models.

Table 6. Variable Significance, NASA/JPL Models, Group 2

Cases < 500 Man-Months of Effort - Variable Significance												
NASA/JPL Model Number	7	7	8	8	9	9	10	10	11	11	12	12
Standard (S), with Lines of Code (L)	S	L	S	L	S	L	S	L	S	L	S	L
MissDSP	--	--	--	4	5	1	5	3	--	--	--	--
ComContrl	1	1	2	--	4	4	4	--	--	--	--	--
DataCC	--	--	4	2	1	5	***1***	2	--	--	--	--
DataP	3	--	3	--	3	3	3	4	--	--	--	--
SimI	--	--	5	1	2	2	2	1	--	--	--	--
AvMon	--	--	--	--	--	--	--	5	--	--	--	--
RELYL	--	--	--	--	--	--	--	--	--	--	--	--
CPLXL	--	--	--	--	--	--	--	--	--	--	--	--
TURNH	--	--	--	--	--	--	--	--	--	3	--	1
ACAPL	--	--	--	--	--	--	--	--	--	--	--	--
VIRTH	--	--	--	--	--	--	--	--	--	--	--	--
LEXPN	--	--	--	--	--	--	--	--	1	--	2	--
PCAPVL	--	2	1	--	--	--	--	--	4	***1***	3	--
STORH	2	5	--	--	--	--	--	--	3	--	4	4
EAF	5	4	--	3	--	--	--	--	5	2	--	--
VEXPL	--	--	--	--	--	--	--	--	--	--	--	--
DATAL	--	--	--	--	--	--	--	--	2	4	1	5
TOOLN	4	3	--	--	--	--	--	--	--	--	--	--
TIMEH	--	--	--	5	--	--	--	--	--	--	--	3
MODPH	--	--	--	--	--	--	--	--	--	5	--	2
SCEDVH	--	--	--	--	--	--	--	--	--	--	5	--
LOC	--	X	--	X	--	X	--	X	--	X	--	X

Table 7 indicates that the simulation variable has the largest magnitude when only the project descriptive variables are used in the models of this group. The high main storage constraint variable consistently has one of the largest magnitudes across the models that include the COCOMO factor-level variables. Only in two instances were any of the top five variables in terms of magnitude significant in the models of this group: data capture/collection and low programmer capability.

Table 7. Variable Magnitude, NASA/JPL Models, Group 2

Cases < 500 Man-Months of Effort - Variable Magnitude												
NASA/JPL Model Number	7	7	8	8	9	9	10	10	11	11	12	12
Standard (S), with Lines of Code (L)	S	L	S	L	S	L	S	L	S	L	S	L
MissDSP	--	--	--	5	5	1	5	2	--	--	--	--
ComContrl	3	5	3	--	4	5	4	--	--	--	--	--
DataP	--	--	5	--	3	3	3	4	--	--	--	--
Siml	--	--	--	1	2	2	2	1	--	--	--	--
DataCC	--	--	4	3	1	4	***1***	3	--	--	--	--
AvMon	--	--	--	--	--	--	--	5	--	--	--	--
RELYL	--	--	--	--	--	--	--	--	--	--	--	--
CPLXL	--	--	--	--	--	--	--	--	--	--	--	--
STORH	1	1	1	--	--	--	--	--	1	--	4	1
TURNH	--	--	--	--	--	--	--	--	--	--	--	--
VIRTH	--	--	--	--	--	--	--	--	--	--	--	--
TURNH	--	--	--	--	--	--	--	--	--	3	--	3
PCAPVL	5	2	2	--	--	--	--	--	4	***1***	1	5
LEXPN	--	--	--	--	--	--	--	--	2	--	2	--
VEXPL	--	--	--	--	--	--	--	--	--	--	--	--
EAF	2	3	--	2	--	--	--	--	5	2	--	--
VEXPL	--	--	--	--	--	--	--	--	--	--	--	--
DATAL	--	--	--	--	--	--	--	--	3	5	3	--
TOOLN	4	4	--	--	--	--	--	--	--	--	--	--
TIMEH	--	--	--	4	--	--	--	--	--	--	--	4
MODPH	--	--	--	--	--	--	--	--	--	4	5	2
ACAPL	--	--	--	--	--	--	--	--	--	--	--	--

NASA/JPL Model Group 3.

The models in this group depicted in Table 8 have reasonable R^2 values, but none of the models is significant. However, the MAPE values for these models are some of the best scores of all the models tested. This is noted with the caveat that these models were built on a very small number of cases and tested on even a smaller reserve set of cases. The models in this group tend to under-predict effort on average as a whole. Only one of the models over-predicts on average. The natural log transformation of the dependent variable in this group of models had mixed results in regard to MAPE scores. One model improved, one was degraded, and the other saw no change to MAPE with this transformation. The models that meet the OLS assumptions in this group tend to have the

lower p-values, but their R^2 values are no higher. Two of the models meeting these assumptions have some of the worst MAPE scores, but one of them has the best MAPE score. Adding lines of code created models with perfect relationships with effort in several of the models in this group. This is an infeasible solution, so these comparisons are not available.

Table 8. Performance Summary, NASA/JPL Models, Group 3

Cases > 500 Man-Months of Effort											
	Project Characteristic Variables	COCOMO Factor-Level Variables	Ln (Effort)	ME	MAE	MPE	MAPE	MSE	R^2	P-value	Meets OLS Assumptions
Model # 13	X	X		530.50	896.50	7.30	52.20	1,085,142.50	0.89	0.61	No
With LOC	X	X		--	--	--	--	--	--	--	--
Model # 14	X	X	X	530.50	896.50	7.30	52.20	1,085,142.40	0.96	0.377	No
With LOC	X	X	X	--	--	--	--	--	--	--	--
Model # 15	X			499.80	927.20	3.51	55.95	1,109,499.88	0.83	0.22	No
With LOC	X			645.24	662.16	26.20	28.28	854,790.27	0.99	0.02	Yes
Model # 16	X		X	480.17	946.83	1.10	58.36	1,127,048.23	0.90	0.1	Yes
With LOC	X		X	560.93	812.57	13.18	44.05	974,915.85	0.98	0.05	Yes
Model # 17		X		-233.70	1,452.70	-78.06	128.86	2,164,952.98	0.89	0.61	No
With LOC		X		--	--	--	--	--	--	--	--
Model # 18		X	X	18.64	1,200.37	-47.10	97.90	1,441,223.84	0.96	0.377	No
With LOC		X	X	--	--	--	--	--	--	--	--

According to Table 9, the mission design/sequencing/planning variable tends to have the lowest p-values across the model types. For the models that included the COCOMO factor-level variables, low required software reliability was consistently significant at a low p-value across the models. Only one variable in one model was statistically significant with a p-value under 0.05 in all the models in this group.

Table 9. Variable Significance, NASA/JPL Models, Group 3

Cases > 500 Man-Months of Effort - Variable Significance												
NASA/JPL Model Number	13	13	14	14	15	15	16	16	17	17	18	18
Standard (S), with Lines of Code (L)	S	L	S	L	S	L	S	L	S	L	S	L
MissDSP	1	N	1	N	1	4	***1***	1	--	N	--	N
ComContrl	4	N	4	N	2	3	2	3	--	N	--	N
DataCC	5	N	5	N	4	1	4	2	--	N	--	N
DataP	--	N	--	N	5	5	5	5	--	N	--	N
SimI	3	N	2	N	3	2	3	4	--	N	--	N
AvMon	--	N	--	N	--	--	--	--	--	N	--	N
RELYL	2	N	3	N	--	--	--	--	1	N	1	N
CPLXL	--	N	--	N	--	--	--	--	--	N	4	N
TURNH	--	N	--	N	--	--	--	--	4	N	--	N
ACAPL	--	N	--	N	--	--	--	--	2	N	2	N
VIRTH	--	N	--	N	--	--	--	--	3	N	5	N
LEXPN	--	N	--	N	--	--	--	--	--	N	--	N
PCAPVL	--	N	--	N	--	--	--	--	--	N	--	N
STORH	--	N	--	N	--	--	--	--	5	N	3	N
EAF	--	N	--	N	--	--	--	--	--	N	--	N
VEXPL	--	N	--	N	--	--	--	--	--	N	--	N
DATAL	--	N	--	N	--	--	--	--	--	N	--	N
TOOLN	--	N	--	N	--	--	--	--	--	N	--	N
TIMEH	--	N	--	N	--	--	--	--	--	N	--	N
MODPH	--	N	--	N	--	--	--	--	--	N	--	N
SCEDVH	--	N	--	N	--	--	--	--	--	N	--	N
LOC	--	N	--	N	--	X	--	O	--	N	--	N

Table 10 shows that the mission design/sequencing/planning variable is also a leader in terms of magnitude for those models that include the project descriptive variables. Across the models in this group that contain the COCOMO factor-level variables, the low required software reliability variable tends to have one of the larger magnitudes.

Table 10. Variable Magnitude, NASA/JPL Models, Group 3

Cases > 500 Man-Months of Effort - Variable Magnitude												
NASA/JPL Model Number	13	13	14	14	15	15	16	16	17	17	18	18
Standard (S), with Lines of Code (L)	S	L	S	L	S	L	S	L	S	L	S	L
MissDSP	1	N	1	N	1	3	***1***	1	--	N	--	N
ComContrl	4	N	4	N	2	4	2	3	--	N	--	N
DataP	--	N	--	N	5	5	5	5	--	N	--	N
Siml	3	N	2	N	3	1	3	4	--	N	--	N
DataCC	5	N	--	N	4	2	4	2	--	N	--	N
AvMon	--	N	--	N	--	--	--	--	--	N	--	N
RELYL	2	N	3	N	--	--	--	--	1	N	1	N
CPLXL	--	N	--	N	--	--	--	--	--	N	5	N
STORH	--	N	--	N	--	--	--	--	5	N	4	N
TURNH	--	N	--	N	--	--	--	--	--	N	--	N
VIRTH	--	N	--	N	--	--	--	--	3	N	3	N
TURNH	--	N	--	N	--	--	--	--	4	N	--	N
PCAPVL	--	N	--	N	--	--	--	--	--	N	--	N
LEXPN	--	N	--	N	--	--	--	--	--	N	--	N
VEXPL	--	N	--	N	--	--	--	--	--	N	--	N
EAF	--	N	--	N	--	--	--	--	--	N	--	N
VEXPL	--	N	--	N	--	--	--	--	--	N	--	N
DATAL	--	N	--	N	--	--	--	--	--	N	--	N
TOOLN	--	N	--	N	--	--	--	--	--	N	--	N
TIMEH	--	N	5	N	--	--	--	--	--	N	--	N
MODPH	--	N	--	N	--	--	--	--	--	N	--	N
ACAPL	--	N	--	N	--	--	--	--	2	N	2	N

COCOMO Estimates.

While the purpose of this research is not necessarily to improve upon or outperform the COCOMO model per se, the variables are available in the dataset to derive estimates using this model for the NASA/JPL projects. These estimates were calculated and reported as a means of analyzing the models created for this thesis. The five error measures were calculated for the COCOMO models using the same sets of reserved data that the derived models used. However, this comparison is greatly dependent on the accuracy of the lines of code used in the COCOMO model.

Gil Barron provides a version of the COCOMO model online (Barron, 1997:n.pag.). The embedded system option was selected for estimating the NASA/JPL

projects. An embedded project “must operate within (is embedded in) a strongly coupled complex of hardware, software, regulations, and operational procedures… (Boehm, 1981:79).” Software projects developed within this complex are forced to comply and conform to the required specifications with little room for tradeoffs or negotiating easier programming options. This significantly increases the amount of effort required to produce such a project (Boehm, 1981:79).

Two sets of error measures are provided in the tables below representing the predictive qualities of the basic and intermediate versions of the model. The basic version incorporates only the lines of codes as an input parameter. The intermediate version includes the lines of code and a multiplier derived from the 15 environmental factor variables. Tables 11 and 12 show that each of the COCOMO variants estimate fairly consistently between models that include all cases, models that include only cases with low effort, and those models including only cases with higher effort levels. It is also apparent that the intermediate version of the COCOMO model outperforms the basic version. The intermediate COCOMO model predicts the NASA/JPL data fairly well; however, as well as it estimates, it still fails to produce MAPE measures at or below 25 percent.

Table 11. Basic COCOMO Performance, NASA/JPL Data

COCOMO - Basic Model					
	ME	MAE	MPE	MAPE	MSE
All	-372.90	372.90	-127.90	127.90	375,939.80
Small	-268.10	268.10	-139.60	139.60	287,792.80
Large	-897.00	897.00	-69.30	69.30	816,674.50

Table 12. Intermediate COCOMO Performance, NASA/JPL Data

COCOMO - Intermediate Model					
	ME	MAE	MPE	MAPE	MSE
All	-87.35	91.90	-25.41	34.22	31,576.42
Small	-63.55	69.01	-25.59	36.15	22,382.04
Large	-206.33	206.33	-24.86	24.53	77,548.36

COCOMO requires the number of lines of code in order to calculate an effort estimate. The models developed in this thesis avoided using lines of code as an input parameter, so comparing their predictive ability to COCOMO is an inadequate comparison of the predictive capabilities of the non-LOC parameters. A separate model was developed for each size-oriented group using only the COCOMO environmental variables and their actual input values. These models were created using the same methods used for the other models including the methods for handling multicollinearity and singularity. The error measures for these models are provided in the model comparison tables that follow.

Analysis of ISBSG Models

Six different models were developed using the 93 randomly selected cases of the 116 cases from the ISBSG dataset. Models 1 and 2 consider all 93 cases. Models 3 and 4 consider only the cases with under 10,000 person-hours of effort. The final two models only consider the cases with over 10,000 person-hours of effort. Each of the six models was tested on the like cases in the reserved set of 23 cases. Like the NASA/JPL models, the two models estimating the larger projects have some limitations. Only 10 projects in the sample fit this size criterion, and only 4 like projects were in the reserved set. While the outputs from these two models should be met with criticism, it is still useful to compare how the method performs on the few cases that are available.

Three separate tables are provided for each of three groups of models in the following sections. The models in the first group are those derived from all of the cases in the sample. The models in the second set are derived from those cases with less than 10,000 person-hours of effort. The models in the third set were derived from only those cases with over 10,000 person-hours of effort. The first table in each set displays whether or not a natural log transformation of the dependent variable was used, if functional points are included, five error measures, the R^2 value, and a p-value for each model. It also shows whether or not the OLS assumptions were met for each model. The second and third charts in each group are like those for the NASA models. They list the top five variables in regard to p-value and magnitude. Instead of lines of code, the number of functional points was added as a variable to each model. The notation used previously to indicate significance is used again here.

The two functions below represent the basic form of the ISBSG models analyzed. The programs' effort or natural log of the effort is a function of the binomial characteristic variables. Functional point values were added to each model for testing and comparison.

Effort in Person-hours = f(33 potential binomial attribute variables, functional points) (3)

Ln(Effort in Person-hours) = f(33 potential binomial attribute variables, functional points) (4)

ISBSG Model Group 1.

Table 13 indicates that all the models in this group are statistically significant and have fairly strong R^2 values. None of the MAPE scores are within the acceptable range

of 25 percent or less. The models without the natural log transformation tend to under-predict effort, and the models with the transformation tend to over-predict effort. The model without functional points and with the natural log transformation has a much worse MAPE score than the same model without the semi-log transformation. The model with functional points and the natural log transformation of effort does perform better than the same model without this transformation. Only one of the models in this group meets all the OLS assumptions. This model has one of the lowest p-values, the second highest R^2 value and the best MAPE score compared to the other models in the group. Adding functional points to the model improved the MAPE scores in both cases.

Table 13. Performance Summary, ISBSG Models, Group 1

All Cases (All Sizes)									
	Ln (Effort)	ME	MAE	MPE	MAPE	MSE	R^2	P-value	Meets OLS Assumptions
Model # 1		96.22	3,756.39	-80.92	211.26	24,564,128.80	0.51	0.0012	No
With FP		634.10	3,325.32	17.42	205.46	16,289,594.30	0.76	< 0.0001	No
Model # 2	X	-58,954.28	62,671.12	-466.08	536.40	82,557,000,000.00	0.59	< 0.0001	No
With FP	X	-7,910.94	11,155.67	-98.60	159.44	1,868,885,720.00	0.73	< 0.0001	Yes

Table 14 shows that the variable representing the Natural programming language consistently has one of the smallest p-values. All five variables in each of the models in this group are statistically significant except one: public administration.

Table 14. Variable Significance, ISBSG Models, Group 1

All Cases (All Sizes) - Variable Significance				
ISBSG Model Number	1	1	2	2
Standard (S), with Functional Points (F)	S	F	S	F
Natural	***1***	***1***	***1***	***2***
Accounting	***2***	***5***	--	--
New Development	***3***	***2***	***5***	***3***
GUI Interface Application	***4***	***3***	--	***5***
Public Administration	5	--	***2***	***1***
Cobol	--	***4***	--	***4***
Logistics	--	--	***3***	--
Access	--	--	***4***	--
Communication	--	--	--	--
Unix Script	--	--	--	--
Telecommunications	--	--	--	--
Engineering	--	--	--	--
Sales and Marketing	--	--	--	--
Sales	--	--	--	--
Manufacturing	--	--	--	--
Transportation/Production System	--	--	--	--
Process Modeling	--	--	--	--
FPs	--	X	--	X

According to Table 15, the variable indicating the Natural programming language has the largest magnitude in all the models in this group. The C programming language and public administration application variables are consistently within to top 5 variables in terms of magnitude for the models in this group.

Table 15. Variable Magnitude, ISBSG Models, Group 1

All Cases (All Sizes) - Variable Magnitude				
ISBSG Model Number	1	1	2	2
Standard (S), with Functional Points (F)	S	F	S	F
Natural	***1***	***1***	***1***	***1***
Accounting	***2***	***4***	--	--
C	3	2	4	3
Logistics	4	--	***2***	5
Public Administration	5	5	***3***	***2***
GUI Interface Application	--	***3***	--	--
Access	--	--	***5***	--
Sales and Marketing	--	--	--	***4***
Telecommunications	--	--	--	--
Communication	--	--	--	--
Unix Script	--	--	--	--
Engineering	--	--	--	--
Manufacturing	--	--	--	--
Sales	--	--	--	--
Transportation/Production System	--	--	--	--
Cobol	--	--	--	--
New Development	--	--	--	--
Process Modeling	--	--	--	--

ISBSG Model Group 2.

Table 16 shows that all the models in this group are statistically significant and have average R^2 values. However, none of the models perform within the acceptable range in terms of MAPE. Three of the models over-predict effort on average, and one under-predicts effort on average. The models with the natural log transformations outperform their like models without the transformation. Including functional points in the model improved the MAPE scores. Only one of the models met the assumptions for OLS. This model had one of the lowest p-values, the highest R^2 value, and the lowest MAPE score.

Table 16. Performance Summary, ISBSG Models, Group 2

Cases < 10,000 Man-Hours of Effort									
	Ln (Effort)	ME	MAE	MPE	MAPE	MSE	R^2	P-value	Meets OLS Assumptions
Model # 3		-688.70	1,757.94	-152.12	180.71	7,044,933.51	0.48	0.0046	No
With FP		-675.90	1,607.04	-125.02	149.02	5,784,743.97	0.51	0.0021	No
Model # 4	X	-60.67	1,917.77	-91.80	151.39	7,629,685.34	0.59	< 0.0001	No
With FP	X	92.06	1,400.82	-56.96	105.38	3,616,263.90	0.64	< 0.0001	Yes

Table 17 indicates that not one variable consistently has the lowest p-value in the models of this group. The engineering and sales and marketing variables are among the five variables with the smallest p-values for each of the models in this group.

Table 17. Variable Significance, ISBSG Models, Group 2

Cases < 10,000 Man-Hours of Effort - Variable Significance				
ISBSG Model Number	3	3	4	4
Standard (S), with Functional Points (F)	S	F	S	F
Natural	--	--	--	--
Accounting	--	--	--	--
New Development	--	--	--	--
GUI Interface Application	--	--	--	--
Public Administration	--	--	--	--
Cobol	--	--	--	--
Logistics	--	--	--	--
Access	--	--	--	--
Communication	***1***	***2***	***4***	--
Unix Script	***2***	***1***	--	***4***
Telecommunications	***3***	***3***	--	--
Engineering	***4***	***4***	***1***	***1***
Sales and Marketing	***5***	***5***	***2***	***3***
Sales	--	--	***3***	***5***
Manufacturing	--	--	***5***	***2***
Transportation/Production System	--	--	--	--
Process Modeling	--	--	--	--
FPs	--	O	--	X

Table 18 shows that the sales and marketing variable appears to have the largest magnitude on average between the models in this group. The engineering variable is among the top five variables in terms of magnitude for all five models.

Table 18. Variable Magnitude, ISBSG Models, Group 2

Cases < 10,000 Man-Hours of Effort - Variable Magnitude				
ISBSG Model Number	3	3	4	4
Standard (S), with Functional Points (F)	S	F	S	F
Natural	--	--	--	--
Accounting	--	--	--	--
C	--	--	--	--
Logistics	--	--	--	4
Public Administration	--	--	--	--
GUI Interface Application	--	--	--	--
Access	--	--	--	--
Sales and Marketing	***1***	***1***	***2***	***3***
Telecommunications	***2***	***2***	5	--
Communication	***3***	***4***	--	--
Unix Script	***4***	***3***	--	--
Engineering	***5***	***5***	***1***	***1***
Manufacturing	--	--	***3***	***2***
Sales	--	--	***4***	***5***
Transportation/Production System	--	--	--	--
Cobol	--	--	--	--
New Development	--	--	--	--
Process Modeling	--	--	--	--

ISBSG Model Group 3.

None of the models in this group, depicted in Table 19, are statistically significant, yet they all have very large R^2 values. All the models in this group meet the OLS assumptions, and they all have some of the better MAPE values recorded in this thesis. However, these MAPE values are still not low enough for the models to be acceptable. The models with the natural log transformation tend to do a little better in regard to MAPE than the standard models. All the models in this group tend to over-predict effort on average. Adding functional points to each of the models improved their MAPE scores.

Table 19. Performance Summary, ISBSG Models, Group 3

Cases > 10,000 Man-Hours of Effort									
	Ln (Effort)	**ME**	**MAE**	**MPE**	**MAPE**	**MSE**	**R^2**	**P-value**	**Meets OLS Assumptions**
Model # 5		-6,018.64	10,691.36	-37.36	84.07	148,061,584.00	0.78	0.57	Yes
With FP		-5,695.49	10,164.53	-32.55	77.22	171,536,121.00	0.97	0.111	Yes
Model # 6	X	-7,427.64	8,350.63	-52.06	61.29	130,123,222.00	0.72	0.68	Yes
With FP	X	-6,814.59	8,082.98	-44.67	57.34	158,940,732.00	0.99	0.046	Yes

According to Table 20, none of the variables in the models of this group consistently has the lowest p-value. The Transportation/Production System variable has the lowest p-value on average and is consistently one of the five lowest p-values among the models. Only two variables in one model (New Development and Transportation/Production System) are statistically significant.

Table 20. Variable Significance, ISBSG Models, Group 3

Cases > 10,000 Man-Hours Effort - Variable Significance				
ISBSG Model Number	5	5	6	6
Standard (S), with Functional Points (F)	S	F	S	F
Natural	--	--	--	--
Accounting	3	--	3	--
New Development	4	2	5	***2***
GUI Interface Application	--	1	--	4
Public Administration	--	--	--	--
Cobol	2	--	2	5
Logistics	--	--	--	--
Access	--	--	--	--
Communication	--	--	--	--
Unix Script	--	--	--	--
Telecommunications	--	--	--	--
Engineering	5	5	4	--
Sales and Marketing	--	--	--	--
Sales	--	--	--	--
Manufacturing	--	--	--	--
Transportation/Production System	1	3	1	***1***
Process Modeling	--	4	--	3
FPs	--	O	--	X

Table 21 shows that the transportation production system variable consistently has one of the largest magnitudes in the models of this group. The engineering variable is

among the top five variables in regard to magnitude for all of the models in this group. Only two of the large magnitude variables in one model are statistically significant: transportation/production system and new development.

Table 21. Variable Magnitude, ISBSG Models, Group 3

Cases > 10,000 Man-Hours of Effort - Variable Magnitude				
ISBSG Model Number	5	5	6	6
Standard (S), with Functional Points (F)	S	F	S	F
Natural	--	--	--	--
Accounting	1	--	1	--
C	--	--	--	--
Logistics	--	--	--	--
Public Administration	--	--	--	--
GUI Interface Application	--	3	--	1
Access	--	--	--	--
Sales and Marketing	--	--	--	--
Telecommunications	--	--	--	--
Communication	--	--	--	--
Unix Script	--	--	--	--
Engineering	5	5	4	5
Manufacturing	--	--	--	--
Sales	--	--	--	--
Transportation/Production System	2	2	2	***2***
Cobol	3	--	3	--
New Development	4	1	5	***3***
Process Modeling	--	4	--	4

Comparison of Performance

Table 22 summarizes the five error measures for all of the estimates derived from the models that included both the cases with small and large effort values. There is a mix of models that over-predict and under-predict on average. None of the models have an acceptable MAPE score of 25 percent or less. All the models, except one, that included the natural log transformation of the effort had better MAPE scores than the like models without this transformation. Adding lines of code or functional points also improved the MAPE scores for each model. The intermediate COCOMO model had the best MAPE

score, but the basic COCOMO model was outperformed by some of the other models. The COCOMO regression has a better MAPE score than Model 3, the model with only the project descriptive variables and no semi-log transformation or line of code added. However, the COCOMO regression underestimates effort on average almost six times the amount that Model 3 overestimates effort on average. The COCOMO regression with the semi-log transformation has a slightly better MAPE than Model 5, the model with a semi-log transformation, the project descriptive variables, and no lines of code. Both the transformed COCOMO regression and Model 4 underestimates effort on average, but the transformed COCOMO regression underestimates by a little less. It appears that the respective categorical models predict the effort for the NASA/JPL projects and the ISBSG projects with about the same level of accuracy according to MAPE scores.

Table 22. Performance Comparison, All-Size Models

All Cases (All Sizes)		ME	MAE	MPE	MAPE	MSE
NASA	**Model # 1**	105.20	438.60	-430.10	526.50	612,840.10
NASA	**With LOC**	-167.10	261.20	-77.20	149.00	153,323.30
NASA	**Model # 2**	255.15	323.74	-64.84	124.00	516,306.53
NASA	**With LOC**	-56.12	124.17	-22.66	42.73	53,949.34
NASA	**Model # 3**	-51.83	512.65	-650.03	675.99	565,643.63
NASA	**With LOC**	-29.11	216.92	-62.01	159.93	121,927.60
NASA	**Model # 4**	207.42	362.68	-162.99	218.82	488,749.93
NASA	**With LOC**	102.63	151.23	-48.27	78.06	91,386.56
NASA	**Model # 5**	53.96	391.60	-203.33	255.04	437,823.34
NASA	**With LOC**	-115.37	184.03	-4.50	118.68	101,695.45
NASA	**Model # 6**	130.73	383.25	-79.16	115.15	566,246.48
NASA	**With LOC**	-28.56	240.92	-61.28	72.80	251,976.92
COCOMO	**Basic**	-372.90	372.90	-127.90	127.90	375,939.80
COCOMO	**Intermediate**	-87.35	91.90	-25.41	34.22	31,576.42
COCOMO	**Regression**	307.32	649.07	-102.46	277.39	1,548,074.79
COCOMO	**Regression (Ln Trans)**	169.49	396.85	-92.29	142.65	592,944.43
ISBSG	**Model # 1**	96.22	3,756.39	-80.92	211.26	24,564,128.80
ISBSG	**With FP**	634.10	3,325.32	17.42	205.46	16,289,594.30
ISBSG	**Model # 2**	-58,954.28	62,671.12	-466.08	536.40	82,557,000,000.00
ISBSG	**With FP**	-7,910.94	11,155.67	-98.60	159.44	1,868,885,720.00

Table 23 summarizes the error measures for the models developed using the cases with the smaller efforts. It appears that more of these models have a tendency to over-predict on average than those that under-predict. Some of the very large MAPE scores in Table 22 are not repeated in these models. However, the average magnitude of the MAPE values does not appear to be much different. None of the MAPE scores in the table are within the acceptable rage. The natural log transformation improved MAPE scores for the standard NASA/JPL models, but the models with the lines of coded added to the transformed models did not show improvement. The natural log transformation did improve all the ISBSG models. Adding lines of code or functional points to the models improved each MAPE score. The intermediate COCOMO model had the best MAPE score of all models presented, but the basic version was outperformed by some of the other NASA models. The COCOMO regression model has a much worse MAPE score than Model 9, the model with only the project descriptive variables and no lines of code parameter. Both the COCOMO regression model and Model 9 overestimate effort on average, but Model 9 overestimate by much less. The COCOMO regression model with the semi-log transformation has a better MAPE score than Model 10, the model with the project descriptive variables, a semi-log transformation, and no lines of code parameter. Both the transformed COCOMO regression model and Model 10 underestimate effort, but Model 10 underestimates effort by less. The table also shows that there were several smaller MAPE scores for the NASA/JPL models than the ISBSG models.

Table 23. Performance Comparison, Lower-Effort Models

Small Cases		ME	MAE	MPE	MAPE	MSE
NASA	Model # 7	-31.78	99.79	-140.85	165.35	12,890.96
NASA	With LOC	-42.64	55.31	-30.50	61.84	6,302.53
NASA	Model # 8	18.27	79.51	-58.29	93.22	11,746.37
NASA	With LOC	-185.83	222.90	-59.70	84.13	332,143.78
NASA	Model # 9	-35.41	125.97	-200.09	228.25	20,126.42
NASA	With LOC	-35.47	64.80	-26.55	65.59	8,280.10
NASA	Model # 10	11.37	110.97	-109.66	149.47	19,932.63
NASA	With LOC	-181.60	214.52	-69.46	83.83	322,189.42
NASA	Model # 11	-88.36	114.38	-191.04	203.14	16,408.68
NASA	With LOC	-61.69	69.02	-63.78	66.69	10,622.82
NASA	Model # 12	-77.41	106.92	-85.39	99.12	24,828.50
NASA	With LOC	-357.22	394.88	-101.67	120.98	1,231,686.98
COCOMO	Basic	-268.10	268.10	-139.60	139.60	287,792.80
COCOMO	Intermediate	-63.55	69.01	-25.59	36.15	22,382.04
COCOMO	Regression	-258.80	273.80	-688.16	694.22	102,421.04
COCOMO	Regression (Ln Trans)	41.92	97.89	-48.39	96.66	18,728.44
ISBSG	Model # 3	-688.70	1,757.94	-152.12	180.71	7,044,933.51
ISBSG	With FP	-675.90	1,607.04	-125.02	149.02	5,784,743.97
ISBSG	Model # 4	-60.67	1,917.77	-91.80	151.39	7,629,685.34
ISBSG	With FP	92.06	1,400.82	-56.96	105.38	3,616,263.90

Table 24 summarizes the error measures for the models developed for the cases with the larger effort levels. The NASA/JPL models appear to under-predict on average, and the ISBSG models tend to over-predict on average for these cases. The magnitude of the error measures are larger for the ISBSG models compared to the NASA/JPL models for the most part. Only Intermediate COCOMO has an acceptable MAPE score. This is the only MAPE performance below 25 percent in this thesis. Several of the models outperform the Basic COCOMO MAPE score. Adding lines of code to these models was often problematic. The results were unfeasible for several of the models after including lines of code, so their results were not included. Adding lines of code improved MAPE scores when they could be effectively added to the models. The natural log transformation of the dependent variable did not have the degree of effect on model performance as in the other groups. In many cases there was little to no improvement in the models with this transformation compared to the models without this transformation.

The COCOMO regression models, one with and the other without the semi-log transformation, have slightly better MAPE scores than Models 15 and 16. Models 15 and 16 use only the project descriptive variables and no lines of code. Model 16 uses the semi-log transformation of effort. The COCOMO regression models overestimate effort on average, and Models 15 and 16 underestimate effort on average.

Table 24. Performance Comparison, Higher-Effort Models

Large Cases		ME	MAE	MPE	MAPE	MSE
NASA	**Model # 13**	530.50	896.50	7.30	52.20	1,085,142.50
NASA	**With LOC**	--	--	--	--	--
NASA	**Model # 14**	530.50	896.50	7.30	52.20	1,085,142.40
NASA	**With LOC**	--	--	--	--	--
NASA	**Model # 15**	499.80	927.20	3.51	55.95	1,109,499.88
NASA	**With LOC**	645.24	662.16	26.20	28.28	854,790.27
NASA	**Model # 16**	480.17	946.83	1.10	58.36	1,127,048.23
NASA	**With LOC**	560.93	812.57	13.18	44.05	974,915.85
NASA	**Model # 17**	-233.70	1,452.70	-78.06	128.86	2,164,952.98
NASA	**With LOC**	--	--	--	--	--
NASA	**Model # 18**	18.64	1,200.37	-47.10	97.90	1,441,223.84
NASA	**With LOC**	--	--	--	--	--
COCOMO	**Basic**	-897.00	897.00	-69.30	69.30	816,674.50
COCOMO	**Intermediate**	-206.33	206.33	-24.86	24.53	77,548.36
COCOMO	**Regression**	-326.96	326.96	-37.34	37.34	173,722.56
COCOMO	**Regression (Ln Trans)**	-297.44	297.44	-27.41	27.41	93,818.68
ISBSG	**Model # 5**	-6,018.64	10,691.36	-37.36	84.07	148,061,584.00
ISBSG	**With FP**	-5,695.49	10,164.53	-32.55	77.22	171,536,121.00
ISBSG	**Model # 6**	-7,427.64	8,350.63	-52.06	61.29	130,123,222.00
ISBSG	**With FP**	-6,814.59	8,082.98	-44.67	57.34	158,940,732.00

V. Conclusion and Recommendations

Review

Software spending is increasing for high-tech development organizations. The impact that these growing costs have on organizations' budgets necessitate the use of models that effectively estimate the costs of new software development endeavors. There is room for improvement in today's software cost models. Size variables such as lines of code and functional points are either not available or are difficult to estimate in the early stages of a new project. Models that depend on these variables are subject to the error involved in misestimating these vital inputs. Also, several of the independent variables in current models require a certain amount of software engineering expertise in order to properly derive their values. Cost estimators are dependent on subject matter experts to obtain these value estimates. This increases the cost of using these models, and if the experts are from the organization developing the project, a conflict of interest exists.

While these software-engineering-related variables may be significantly related to development costs, they do not allow a customer to see what a specific set of attributes or functions cost in a particular project. These attributes and functions are what a customer is buying, and it would be more meaningful if cost estimates could be related to their values. Another related obstacle to fully understanding software cost estimates is that several of the commercial models in use are proprietary and not all the internal calculations of the models are available for inspection and analysis. It is not easy to understand and explain how an estimate was derived using these models, and it is very difficult to justify the development costs to a customer or to those funding the project.

The purpose of this thesis is to partially remedy these issues by determining if a hedonic approach, using ordinary least squares regression and nominal attribute variables, can be used to develop a model that adequately estimates NASA and JPL software development costs. The goal is that the attribute variables used will be more meaningful to the customer, cheaper to obtain, and more readily available.

Ordinary least squares regression was the tool used to develop the models for analysis. Two different databases were used. The primary database was a set of 60 software projects developed by NASA or JPL. The second set of data was derived from the International Software Benchmarking and Standards Group database. The analysis compared several cost models containing multiple combinations of available categorical variables from the respective database used. Also, the input variables to operate the COCOMO model were available for the NASA/JPL data. Estimates were created using COCOMO as another means of comparing model performance. The analysis compared the predictive qualities of the models developed using five error measures: mean error, mean absolute error, mean percentage error, mean absolute percentage error, and mean squared error. The MAPE, often referred to as MMRE in the literature, was the primary metric for comparing models.

Conclusions/Recommendations

While the software attribute and characteristic variables and ordinary least squares regression method created models that were simple, easy to understand, and cheap to use, the overall performance of the models was unsatisfactory. None of the models estimated effort for the respective cases in the reserved data set in a manner that produced a MAPE

score of 25 percent or less. Only Model 15 with lines of code added as an independent variable produced a MAPE less than 30 percent. However, since there is a very small amount of data available, and even a smaller amount of the large programs in this set, it is unlikely that a similar outcome would be reproduced in the future. When used in models together, very few of the project descriptive variables in the NASA/JPL data set were better predictors of effort than the modified COCOMO variables in regard to p-value. The intermediate version of COCOMO outperformed all of the models developed using the NASA/JPL data except the one large project, but only when used on the larger projects alone was a MAPE scores under 25 percent reached. This was the only instance that any model met this performance standard.

Size variables such as lines of code or functional points were purposely left out of the primary models due to the associated problems with using them. However, in almost every case, MAPE values were greatly improved when these size variables were added to each model developed. It appears that even though there are problematic issues with using current sizing variables, the categorical variables tested in this research are not an adequate proxy for size when using MAPE as the metric for performance.

In most of the models, the natural log transformation of the effort values tended to improve the predictive capabilities of the models. This was expected since many of the methods and models discussed in previous research utilized this transformation. Variance in the residuals tends to increase as effort increases, and this transformation helps to correct this issue somewhat.

When comparing the significant variables and model performance between the NASA/JPL and ISBSG models, there is little evidence that the more organizationally

specific variables associated with the NASA/JPL data are any more accurate than the more general variables and data available in the ISBSG dataset. However, this is not a perfect comparison since the developed models used different databases and variables, and could not be tested on the projects from the other set. The NASA projects are probably more complex increasing the difficulty of the estimates, and the general variables used in the ISBSG models make it more difficult to account for variations due to more specific attribute differences. It would be very difficult to build one model that could be used to estimate projects from both databases using the methods in this research.

The NASA/JPL models using cases of all sizes performed worse than the models specifically for large or small cases when comparing MAPE values. The models estimating the larger projects had better MAPE scores than the models estimating the smaller projects. The small project models tended to over-estimate effort while the large project models tended to under-estimate effort. The ISBSG models developed using cases of all sizes also performed worse than the models developed specifically for large or small cases. The models for large and small cases tended to over-predict in both cases on average.

While MAPE is commonly used as a metric for evaluating and comparing software cost models, a ME measure may be more appropriate and more meaningful to management. Managers are probably more concerned with the actual effort levels that a particular model either overestimates or underestimates rather than a relative measure. Also, managers should find models that have a tendency to overestimate more useful than models that underestimate since it is better to find out a project will cost less than

expected than to find out it will exceed what was budgeted. The tendency to over or underestimate is revealed by the ME measure.

Analysis of the COCOMO regression models and the NASA/JPL models using only the project specific variables revealed mixed results as to which models are preferred when comparing MAPE scores. However, when considering ME scores, and preferring models that overestimate to those that underestimate, the models based on the project descriptive variables alone are favored to the COCOMO regression models for the small projects under 500 person-months. The standard model using the project descriptive variables overestimated effort by 223.39 person-months less than the COCOMO based regression model on average. This equates to 86.3 percent less error on average. The project descriptive model also had a better MAPE score. The project descriptive model with the semi-log transformation had a 1.5 times larger MAPE than the transformed COCOMO based model, but it underestimated effort by 30.55 person-months less on average. This is 72.9 percent less error on average. However, for the few cases that fell above the 500 person-month threshold, the COCOMO regressions were the better choice. These results indicate that managers would be better off using the simple models based on project descriptive variables for smaller projects rather than more complex and expensive methods. The more complex and expensive models requiring software expertise should be reserved for the few larger projects.

Limitations

When the large and small cases were analyzed graphically together in any model, a number of larger cases were found to lie away form the bulk of the other cases. Spyros

Makridakis et al. refer to this as the "King Kong effect (Makridakis, 1998: 197)." This observation led to separating the cases into large and small effort categories for analysis on their own. However, when split at the specified threshold, the effect was still present even in the separated groups but not to the same degree. This issue was related to some of the singularity issues that had to be resolved during model development since a few of the variables were specific to the large projects.

Another limitation encountered was the limited data available for analysis. It was very difficult to find or create a database of software projects that had actual costs or effort values for specific projects. It was even more difficult to find this data with a good amount of descriptive information that could be turned into variables for modeling the costs. Another variable that was not obtained, but could have been useful for analysis and comparison, was the initial effort estimates at the beginning of the projects.

Previous research conducted using the ISBSG database, and the information available, offered some assistance with selecting variables for the models using the ISBSG cases as a source. However, the nature of the descriptive information in the NASA/JPL source data was not as helpful. There was no guidance from previous research either. The method used to form categories was fairly subjective and dependent on the researcher's best judgment. The goal was to find meaningful descriptive variables that were not so specific that they were only related to one or two cases and general enough that the model could possibly be applied to future projects. Even then, the variables used could be too specific to be used in the future.

Future Research

Software cost estimation is quite complex, and there are many questions yet to be answered. Several questions stem from this research that could possibly be considered in future studies. If initial estimates for completed software projects could be obtained with final costs and a good number of descriptive variables, it would be interesting to see how the models developed according the methods in this research would compare to the initial estimate obtained from other models.

Also, there are other categorical regression techniques that could be tested to determine if the models created are any more accurate in their predictions. However, the simplicity and ease of use of the OLS method would be lost with these more complex techniques.

There are different methods of analyzing the predictive qualities of a model than what were used in this model. While MAPE, or MMRE, is the standard measure used in software cost research, there may be a more appropriate measure. Different metrics may regard the models and methods derived in this research more favorably.

Ian Sommerville states, “Project cost estimates are often self-fulfilling. The estimate is used to define the project budget, and the product is adjusted so that the budget figure is realized (Sommerville, 2004:620).” If this is true, this could have serious implications in analyzing how well models really estimate software costs. Current methods may only prove to be good predictors of cost because program management or developers made the project fit estimates derived from these methods and models. It may be difficult to remove this bias from analysis, but it could prove to be beneficial research to study the effects of this phenomenon in regard to government projects.

Appendix A. NASA/JPL Model Variables

This table lists the potential independent variables for the NASA/JPL Models. The models did not include every variable after multicollinearity and singularity issues were resolved.

Table 25. NASA/JPL Model Variables

NASA/JPL Model Variables	
Project Descriptive Variables	
Abbreviation	**Variable**
AvMon	Avionics Monitoring
ComContrl	Command and Control
DataCC	Data Collection, Data Capture
DataP	Data Processing
MissDSP	Mission Design, Mission Sequencing, and Mission Planning
Simul	Simulation
COCOMO Factor-Level and Other Variables	
ACAPL	Analyst Capability - Low
AEXPL	Applications Experience - Low
TURNH	Computer Turnaround Time - High
DATAL	Database Size - Low
EAF	Effort Multiplier Based on COCOMO Factors
TIMEH	Execution Time Constraint - High
STORH	Main Storage Constraint - High
MODPH	Modern Programming Practices - High
CPLXL	Product Complexity - Low
PCAPVL	Programmer Capability - Very Low
LEXPN	Programming Language Experience - Nominal
SCEDVH	Require Development Schedule - Very High
RELYL	Required Software Reliability - Low
LOC/SLOC	Lines of Code/Source Lines of Code (Used Interchangeably)
TOOLN	Use of Software Tools - Nominal
VEXPL	Virtual Machine Experience - Low
VIRTH	Virtual Machine Volatility - High

Appendix B. ISBSG Model Variables

This table lists the potential independent variables for the ISBSG Models. The models did not include every variable after multicollinearity and singularity issues were resolved.

Table 26. ISBSG Model Variables

ISBSG Model Variables	
Variable (Abbreviation)	**Category**
GUI Interface Application	Application Type
Transaction/Production System	Application Type
Accounting	Business Area
Engineering	Business Area
Logistics	Business Area
Manufacturing	Business Area
Medical and Healthcare	Business Area
Research and Development	Business Area
Sales	Business Area
Sales and Marketing	Business Area
Telecommunications	Business Area
PC	Development Platform
Joint Application Development (JAD)	Development Technology
Process Modeling	Development Technology
Structured	Development Technology
New Development	Development Type
4th Generation Language (4GL)	Language Type
Application Generator (APG)	Language Type
Communication	Organization Type
Electricity, Gas, and Water	Organization Type
Manufacturing 2	Organization Type
Medical and Healthcare 2	Organization Type
Public Administration	Organization Type
Revenue	Organization Type
Wholesale and Retail Trade	Organization Type
Access	Programming Language
C	Programming Language
COBOL	Programming Language
Focus	Programming Language
Natural	Programming Language
PowerBuilder	Programming Language
Programming Language I (PL/I)	Programming Language
Unix Script	Programming Language
Functional Points (FP)	Size Variable

Appendix C. NASA/JPL Model Regression Outputs

Model 1: All Sizes, Project Descriptive Variables, COCOMO Factor-level Variables and EAF

Summary of Fit

RSquare	0.738365
RSquare Adj	0.560827
Root Mean Square Error	436.0371
Mean of Response	415.85
Observations (or Sum Wgts)	48

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	19	15023799	790726	4.1589
Error	28	5323594	190128	Prob > F
C. Total	47	20347393		0.0003

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	159.31979	459.6926	0.35	0.7315	.
DataCC	399.59577	326.5442	1.22	0.2313	4.1011238
DataP	81.069914	240.3749	0.34	0.7384	2.0260035
AvMon	308.28716	299.6225	1.03	0.3123	4.869292
Simul	278.10245	340.3631	0.82	0.4208	4.4555755
MissDSP	755.38319	306.2229	2.47	0.0200	4.1819765
ComContrl	517.03488	313.7797	1.65	0.1106	4.0996333
RELYL	1958.7444	819.1471	2.39	0.0238	3.4556801
CPLXL	1232.5652	682.8029	1.81	0.0818	4.699919
TIMEH	-234.7996	406.4053	-0.58	0.5681	5.1941259
STORH	705.74762	425.4276	1.66	0.1083	4.997625
VIRTH	522.18695	723.8235	0.72	0.4766	2.6982059
TURNH	698.07313	371.2571	1.88	0.0705	7.4759544
ACAPL	-429.2915	250.9856	-1.71	0.0982	3.6376427
AEXPL	-430.0197	267.0713	-1.61	0.1186	4.4314864
PCAPVL	183.77001	350.9694	0.52	0.6047	3.4013475
VEXPL	-532.1479	770.1227	-0.69	0.4953	3.0544258
LEXPN	102.4837	351.7633	0.29	0.7729	6.1691338
MODPH	535.72135	486.6687	1.10	0.2804	9.1093185
EAF	-447.6664	393.6338	-1.14	0.2651	3.5362096

Model 1 with Lines of Code Added

Summary of Fit

RSquare	0.97539
RSquare Adj	0.95716
Root Mean Square Error	136.1853
Mean of Response	415.85
Observations (or Sum Wgts)	48

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	20	19846639	992332	53.5053
Error	27	500753	18546	Prob > F
C. Total	47	20347393		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-84.18724	144.3654	-0.58	0.5646	.
DataCC	29.043822	104.5445	0.28	0.7833	4.3093145
DataP	55.657923	75.09162	0.74	0.4650	2.0268962
AvMon	182.33108	93.90498	1.94	0.0627	4.9032154
Simul	13.553919	107.5623	0.13	0.9007	4.5616897
MissDSP	221.23312	101.2146	2.19	0.0377	4.6835988
ComContrl	38.608179	102.3936	0.38	0.7091	4.4753585
RELYL	-217.626	289.2557	-0.75	0.4583	4.4173352
CPLXL	-1212.212	261.654	-4.63	<.0001	7.0752391
TIMEH	-77.33112	127.3056	-0.61	0.5486	5.2248674
STORH	-22.98721	140.3462	-0.16	0.8711	5.575715
VIRTH	705.97528	226.3552	3.12	0.0043	2.7050638
TURNH	357.76086	117.8577	3.04	0.0053	7.7235918
ACAPL	-76.6668	81.38191	-0.94	0.3545	3.9207115
AEXPL	-91.95239	86.0072	-1.07	0.2945	4.7114137
PCAPVL	-287.8398	113.4508	-2.54	0.0173	3.6434627
VEXPL	-49.78436	242.3814	-0.21	0.8388	3.101665
LEXPN	282.26661	110.4287	2.56	0.0165	6.2326616
MODPH	182.95396	153.5649	1.19	0.2439	9.2980045
EAF	-217.7789	123.7654	-1.76	0.0898	3.5837572
KLOC	6.9307535	0.429793	16.13	<.0001	4.5586368

Model 2: All Sizes, Project Descriptive Variables, COCOMO Factor-level Variables and EAF with Natural Log Transformation of Effort

Summary of Fit

RSquare	0.620327
RSquare Adj	0.362692
Root Mean Square Error	1.105275
Mean of Response	5.118768
Observations (or Sum Wgts)	48

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	19	55.886816	2.94141	2.4078
Error	28	34.205692	1.22163	Prob > F
C. Total	47	90.092508		0.0170

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	4.0992951	1.165237	3.52	0.0015	.
DataCC	1.1775422	0.82773	1.42	0.1659	4.1011238
DataP	-0.461685	0.609307	-0.76	0.4549	2.0260035
AvMon	0.0714916	0.759488	0.09	0.9257	4.869292
Simul	1.0123043	0.862758	1.17	0.2505	4.4555755
MissDSP	0.385827	0.776219	0.50	0.6230	4.1819765
ComContrl	1.3173275	0.795374	1.66	0.1088	4.0996333
RELYL	1.2354486	2.076389	0.59	0.5566	3.4556801
CPLXL	3.755744	1.730781	2.17	0.0386	4.699919
TIMEH	-0.077322	1.030163	-0.08	0.9407	5.1941259
STORH	1.7082526	1.078381	1.58	0.1244	4.997625
VIRTH	-0.57594	1.834761	-0.31	0.7559	2.6982059
TURNH	0.3985014	0.941069	0.42	0.6752	7.4759544
ACAPL	-0.368191	0.636203	-0.58	0.5674	3.6376427
AEXPL	-0.391733	0.676977	-0.58	0.5675	4.4314864
PCAPVL	1.3183623	0.889644	1.48	0.1495	3.4013475
VEXPL	-0.233963	1.952121	-0.12	0.9055	3.0544258
LEXPN	-0.627711	0.891656	-0.70	0.4873	6.1691338
MODPH	0.5223039	1.233617	0.42	0.6752	9.1093185
EAF	0.017998	0.99779	0.02	0.9857	3.5362096

Model 2 with Lines of Code Added

Summary of Fit

RSquare	0.84194
RSquare Adj	0.724858
Root Mean Square Error	0.726229
Mean of Response	5.118768
Observations (or Sum Wgts)	48

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	20	75.852479	3.79262	7.1911
Error	27	14.240028	0.52741	Prob > F
C. Total	47	90.092508		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	3.6038427	0.769851	4.68	<.0001	.
DataCC	0.4235975	0.5575	0.76	0.4539	4.3093145
DataP	-0.513389	0.400438	-1.28	0.2107	2.0268962
AvMon	-0.184785	0.500763	-0.37	0.7150	4.9032154
Simul	0.4740398	0.573593	0.83	0.4158	4.5616897
MissDSP	-0.700983	0.539743	-1.30	0.2050	4.6835988
ComContrl	0.343895	0.54603	0.63	0.5341	4.4753585
RELYL	-3.192711	1.542501	-2.07	0.0482	4.4173352
CPLXL	-1.218529	1.39531	-0.87	0.3902	7.0752391
TIMEH	0.243072	0.678877	0.36	0.7231	5.2248674
STORH	0.2255299	0.748418	0.30	0.7655	5.575715
VIRTH	-0.201995	1.207074	-0.17	0.8683	2.7050638
TURNH	-0.293916	0.628494	-0.47	0.6438	7.7235918
ACAPL	0.3492784	0.433982	0.80	0.4280	3.9207115
AEXPL	0.2961164	0.458647	0.65	0.5240	4.7114137
PCAPVL	0.3587999	0.604994	0.59	0.5581	3.6434627
VEXPL	0.7474793	1.292536	0.58	0.5679	3.101665
LEXPN	-0.261915	0.588878	-0.44	0.6600	6.2326616
MODPH	-0.195455	0.818909	-0.24	0.8132	9.2980045
EAF	0.4857394	0.659998	0.74	0.4681	3.5837572
KLOC	0.0141017	0.002292	6.15	<.0001	4.5586368

Model #3: All Sizes, Project Descriptive Variables Only

Summary of Fit

RSquare	0.205874
RSquare Adj	0.08966
Root Mean Square Error	627.7795
Mean of Response	415.85
Observations (or Sum Wgts)	48

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	4189000	698167	1.7715
Error	41	16158393	394107	Prob > F
C. Total	47	20347393		0.1291

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	135.07481	264.4748	0.51	0.6123	.
DataCC	324.23839	301.67	1.07	0.2887	1.6885573
DataP	-144.3088	271.1357	-0.53	0.5974	1.2435631
AvMon	-92.68356	268.314	-0.35	0.7315	1.8838067
Simul	655.60198	273.294	2.40	0.0211	1.3858361
MissDSP	302.03374	297.4285	1.02	0.3158	1.9032859
ComContrl	388.09982	278.9494	1.39	0.1716	1.5630713

Model #3 with Lines of Code Added

Summary of Fit

RSquare	0.898608
RSquare Adj	0.880865
Root Mean Square Error	227.1044
Mean of Response	415.85
Observations (or Sum Wgts)	48

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	18284337	2612048	50.6442
Error	40	2063056	51576	Prob > F
C. Total	47	20347393		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-291.457	99.0938	-2.94	0.0054	.
DataCC	196.14376	109.4063	1.79	0.0806	1.6970697
DataP	54.204135	98.81788	0.55	0.5864	1.2622015
AvMon	182.37679	98.48053	1.85	0.0714	1.9391601
Simul	405.72304	100.0151	4.06	0.0002	1.418229
MissDSP	236.5724	107.67	2.20	0.0339	1.9058637
ComContrl	155.00208	101.8925	1.52	0.1361	1.5935884
KLOC	5.9399992	0.359314	16.53	<.0001	1.145709

Model #4: All Sizes, Project Descriptive Variables Only with Natural Log Transformation of Effort

Summary of Fit

RSquare	0.28851
RSquare Adj	0.18439
Root Mean Square Error	1.250365
Mean of Response	5.118768
Observations (or Sum Wgts)	48

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	25.992626	4.33210	2.7709
Error	41	64.099882	1.56341	Prob > F
C. Total	47	90.092508		0.0235

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	4.8157916	0.526761	9.14	<.0001	.
DataCC	1.050588	0.600844	1.75	0.0879	1.6885573
DataP	-0.654687	0.540028	-1.21	0.2323	1.2435631
AvMon	-0.621905	0.534408	-1.16	0.2513	1.8838067
Simul	1.4260534	0.544327	2.62	0.0123	1.3858361
MissDSP	0.0114238	0.592396	0.02	0.9847	1.9032859
ComContrl	0.6693493	0.555591	1.20	0.2352	1.5630713

Model #4 with Lines of Code Added

Summary of Fit

RSquare	0.787203
RSquare Adj	0.749963
Root Mean Square Error	0.692305
Mean of Response	5.118768
Observations (or Sum Wgts)	48

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	70.921049	10.1316	21.1389
Error	40	19.171458	0.4793	Prob > F
C. Total	47	90.092508		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	4.054284	0.302078	13.42	<.0001	.
DataCC	0.8218945	0.333514	2.46	0.0181	1.6970697
DataP	-0.300273	0.301236	-1.00	0.3249	1.2622015
AvMon	-0.130826	0.300208	-0.44	0.6653	1.9391601
Simul	0.9799325	0.304886	3.21	0.0026	1.418229
MissDSP	-0.105447	0.328221	-0.32	0.7497	1.9058637
ComContrl	0.2531887	0.310609	0.82	0.4198	1.5935884
KLOC	0.010605	0.001095	9.68	<.0001	1.145709

Model #5: All Sizes, COCOMO Factor-level Variables and EAF Only

Summary of Fit

RSquare	0.676097
RSquare Adj	0.508922
Root Mean Square Error	461.0848
Mean of Response	415.85
Observations (or Sum Wgts)	48

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	16	13756819	859801	4.0442
Error	31	6590574	212599	Prob > F
C. Total	47	20347393		0.0004

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	999.16222	575.334	1.74	0.0924	.
RELYL	1663.7048	668.4484	2.49	0.0184	2.0579311
DATAL	-87.14875	263.1445	-0.33	0.7427	3.8474126
CPLXL	597.26689	602.7471	0.99	0.3294	3.2753324
TIMEH	23.106654	394.5051	0.06	0.9537	4.3770799
STORH	1002.326	369.7691	2.71	0.0108	3.37644
VIRTH	769.78547	941.1395	0.82	0.4196	4.0794639
TURNH	512.29661	373.2319	1.37	0.1797	6.7570948
ACAPL	-293.8217	317.3344	-0.93	0.3616	5.2004655
AEXPL	-250.1889	248.7947	-1.01	0.3224	3.4392399
PCAPVL	249.3991	359.0807	0.69	0.4925	3.1840647
VEXPL	-952.7567	636.8095	-1.50	0.1447	1.8677303
LEXPN	2.5862833	371.2769	0.01	0.9945	6.1461681
MODPH	222.9872	337.2271	0.66	0.5133	3.9115538
TOOLN	-57.19637	385.6594	-0.15	0.8831	7.8704355
SCEDVH	-144.8954	273.6637	-0.53	0.6003	4.161163
EAF	-667.6575	431.1289	-1.55	0.1316	3.7936116

Model #5 with Lines of Code Added

Summary of Fit

RSquare	0.970485
RSquare Adj	0.95376
Root Mean Square Error	141.486
Mean of Response	415.85
Observations (or Sum Wgts)	48

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	17	19746844	1161579	58.0259
Error	30	600548	20018	Prob > F
C. Total	47	20347393		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-12.7395	185.9831	-0.07	0.9458	.
RELYL	-418.355	237.8234	-1.76	0.0888	2.7665526
DATAL	107.71335	81.52908	1.32	0.1964	3.922293
CPLXL	-1320.963	215.6517	-6.13	<.0001	4.4527196
TIMEH	23.248449	121.0557	0.19	0.8490	4.3770799
STORH	-4.483072	127.5225	-0.04	0.9722	4.2648734
VIRTH	607.74711	288.9449	2.10	0.0439	4.0837559
TURNH	247.28591	115.5481	2.14	0.0406	6.8780047
ACAPL	-25.32706	98.60483	-0.26	0.7990	5.3325987
AEXPL	-43.21618	77.27571	-0.56	0.5801	3.523717
PCAPVL	-318.3158	114.9694	-2.77	0.0096	3.466547
VEXPL	-105.8246	201.4483	-0.53	0.6032	1.9849836
LEXPN	275.84731	115.018	2.40	0.0229	6.2643353
MODPH	155.58102	103.553	1.50	0.1434	3.9171005
TOOLN	-26.1405	118.355	-0.22	0.8267	7.8722469
SCEDVH	38.574837	84.64213	0.46	0.6519	4.2275441
EAF	-226.2244	134.7327	-1.68	0.1035	3.9347683
KLOC	7.2365369	0.41834	17.30	<.0001	4.0013811

Model #6: All Sizes, COCOMO Factor-level Variables and EAF Only with Natural Log Transformation of Effort

Summary of Fit

RSquare	0.579569
RSquare Adj	0.362573
Root Mean Square Error	1.105378
Mean of Response	5.118768
Observations (or Sum Wgts)	48

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	16	52.214849	3.26343	2.6709
Error	31	37.877659	1.22186	Prob > F
C. Total	47	90.092508		0.0093

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	6.0327975	1.379272	4.37	0.0001	.
RELYL	1.1189512	1.602499	0.70	0.4902	2.0579311
DATAL	-1.148695	0.630847	-1.82	0.0783	3.8474126
CPLXL	2.3659321	1.444991	1.64	0.1117	3.2753324
TIMEH	0.2555022	0.945764	0.27	0.7888	4.3770799
STORH	1.803111	0.886463	2.03	0.0506	3.37644
VIRTH	1.0802535	2.256233	0.48	0.6355	4.0794639
TURNH	0.5727993	0.894764	0.64	0.5268	6.7570948
ACAPL	-0.455097	0.760759	-0.60	0.5540	5.2004655
AEXPL	-0.288246	0.596446	-0.48	0.6323	3.4392399
PCAPVL	1.3052552	0.860839	1.52	0.1396	3.1840647
VEXPL	-1.188775	1.52665	-0.78	0.4421	1.8677303
LEXPN	-1.101084	0.890078	-1.24	0.2254	6.1461681
MODPH	0.2906135	0.808449	0.36	0.7217	3.9115538
TOOLN	0.2394277	0.924557	0.26	0.7974	7.8704355
SCEDVH	-0.378483	0.656065	-0.58	0.5682	4.161163
EAF	-0.60431	1.033563	-0.58	0.5630	3.7936116

Model #6 with Lines of Code Added

Summary of Fit

RSquare	0.813363
RSquare Adj	0.707602
Root Mean Square Error	0.748657
Mean of Response	5.118768
Observations (or Sum Wgts)	48

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	17	73.277893	4.31046	7.6906
Error	30	16.814615	0.56049	Prob > F
C. Total	47	90.092508		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	4.1352853	0.984108	4.20	0.0002	.
RELYL	-2.785315	1.258415	-2.21	0.0346	2.7665526
DATAL	-0.78329	0.431402	-1.82	0.0794	3.922293
CPLXL	-1.231122	1.141096	-1.08	0.2892	4.4527196
TIMEH	0.2557681	0.640553	0.40	0.6925	4.3770799
STORH	-0.084852	0.674771	-0.13	0.9008	4.2648734
VIRTH	0.7764001	1.528919	0.51	0.6153	4.0837559
TURNH	0.0758528	0.611409	0.12	0.9021	6.8780047
ACAPL	0.0483826	0.521756	0.09	0.9267	5.3325987
AEXPL	0.0998677	0.408896	0.24	0.8087	3.523717
PCAPVL	0.2406795	0.608347	0.40	0.6952	3.466547
VEXPL	0.3993874	1.065941	0.37	0.7105	1.9849836
LEXPN	-0.588667	0.608605	-0.97	0.3412	6.2643353
MODPH	0.1642138	0.547939	0.30	0.7665	3.9171005
TOOLN	0.2976635	0.626262	0.48	0.6380	7.8722469
SCEDVH	-0.034441	0.447874	-0.08	0.9392	4.2275441
EAF	0.2234623	0.712922	0.31	0.7561	3.9347683
KLOC	0.0135699	0.002214	6.13	<.0001	4.0013811

Model #7: Cases with < 500 Person-months Effort, Project Descriptive Variables, COCOMO Factor-level Variables and EAF

Summary of Fit

RSquare	0.34561
RSquare Adj	0.005327
Root Mean Square Error	137.0957
Mean of Response	159.2154
Observations (or Sum Wgts)	39

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	13	248162.81	19089.4	1.0157
Error	25	469880.66	18795.2	Prob > F
C. Total	38	718043.47		0.4669

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	329.61144	178.2249	1.85	0.0763	.
DataCC	96.00959	103.7193	0.93	0.3635	3.2874167
DataP	-93.86712	75.50288	-1.24	0.2253	1.5398557
AvMon	63.742651	92.28045	0.69	0.4961	4.1822472
Simul	63.763359	100.8127	0.63	0.5328	2.3570406
MissDSP	4.2964085	95.50312	0.04	0.9645	3.359593
ComContrl	145.16858	84.51948	1.72	0.0982	2.4168693
TIMEH	-88.90808	119.9767	-0.74	0.4656	2.7492228
STORH	252.68467	189.7019	1.33	0.1949	3.6329795
AEXPL	-40.31582	81.77655	-0.49	0.6263	3.2843381
PCAPVL	114.5713	110.6112	1.04	0.3102	2.3367581
MODPH	-16.30836	167.7652	-0.10	0.9233	5.3755084
TOOLN	-123.8239	106.2479	-1.17	0.2548	4.7432856
EAF	-150.9666	131.6377	-1.15	0.2623	3.2934904

Model #7 with Lines of Code Added

Summary of Fit

RSquare	0.918906
RSquare Adj	0.871601
Root Mean Square Error	49.25657
Mean of Response	159.2154
Observations (or Sum Wgts)	39

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	14	659814.44	47129.6	19.4252
Error	24	58229.03	2426.2	Prob > F
C. Total	38	718043.47		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	115.08312	66.11783	1.74	0.0946	.
DataCC	30.090419	37.60695	0.80	0.4315	3.3480452
DataP	-12.97437	27.82891	-0.47	0.6453	1.6205588
AvMon	17.753136	33.34254	0.53	0.5993	4.2296742
Simul	37.173353	36.27807	1.02	0.3157	2.3645274
MissDSP	-31.702	34.42405	-0.92	0.3663	3.381387
ComContrl	51.961395	31.19836	1.67	0.1088	2.5510704
TIMEH	2.1678756	43.66936	0.05	0.9608	2.8215566
STORH	79.730721	69.43856	1.15	0.2622	3.7708585
AEXPL	-11.70614	29.46315	-0.40	0.6946	3.3026922
PCAPVL	-64.32912	42.04742	-1.53	0.1391	2.6158559
MODPH	14.495058	60.32206	0.24	0.8121	5.3837827
TOOLN	-57.85145	38.50792	-1.50	0.1461	4.8267847
EAF	-60.3116	47.80492	-1.26	0.2192	3.3648081
KLOC	3.6280481	0.27853	13.03	<.0001	1.5549672

Model #8: Cases with < 500 Person-months Effort, Project Descriptive Variables, COCOMO Factor-level Variables and EAF with Natural Log Transformation of Effort

Summary of Fit

RSquare	0.320188
RSquare Adj	-0.03331
Root Mean Square Error	1.041203
Mean of Response	4.639606
Observations (or Sum Wgts)	39

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	13	12.765172	0.98194	0.9058
Error	25	27.102583	1.08410	Prob > F
C. Total	38	39.867755		0.5594

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	4.810492	1.353568	3.55	0.0015	.
DataCC	0.7811223	0.787718	0.99	0.3309	3.2874167
DataP	-0.749441	0.573423	-1.31	0.2031	1.5398557
AvMon	0.2469115	0.700844	0.35	0.7276	4.1822472
Simul	0.6774325	0.765644	0.88	0.3847	2.3570406
MissDSP	-0.034771	0.725319	-0.05	0.9621	3.359593
ComContrl	0.860695	0.641901	1.34	0.1920	2.4168693
TIMEH	-0.304644	0.911189	-0.33	0.7409	2.7492228
STORH	1.1678022	1.440732	0.81	0.4253	3.6329795
AEXPL	-0.056651	0.62107	-0.09	0.9280	3.2843381
PCAPVL	1.1455343	0.84006	1.36	0.1848	2.3367581
MODPH	-0.312379	1.274129	-0.25	0.8083	5.3755084
TOOLN	-0.597603	0.806923	-0.74	0.4658	4.7432856
EAF	-0.216276	0.999751	-0.22	0.8305	3.2934904

Model #8 with Lines of Code Added

Summary of Fit

RSquare	0.800044
RSquare Adj	0.683403
Root Mean Square Error	0.576332
Mean of Response	4.639606
Observations (or Sum Wgts)	39

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	14	31.895950	2.27828	6.8590
Error	24	7.971805	0.33216	Prob > F
C. Total	38	39.867755		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	3.3480252	0.773619	4.33	0.0002	.
DataCC	0.331743	0.440024	0.75	0.4582	3.3480452
DataP	-0.197985	0.325615	-0.61	0.5489	1.6205588
AvMon	-0.066605	0.390128	-0.17	0.8659	4.2296742
Simul	0.4961651	0.424476	1.17	0.2539	2.3645274
MissDSP	-0.280177	0.402782	-0.70	0.4934	3.381387
ComContrl	0.2252897	0.36504	0.62	0.5429	2.5510704
TIMEH	0.3162327	0.510958	0.62	0.5418	2.8215566
STORH	-0.011247	0.812474	-0.01	0.9891	3.7708585
AEXPL	0.1383851	0.344737	0.40	0.6917	3.3026922
PCAPVL	-0.074053	0.491981	-0.15	0.8816	2.6158559
MODPH	-0.102388	0.705805	-0.15	0.8859	5.3837827
TOOLN	-0.14786	0.450566	-0.33	0.7456	4.8267847
EAF	0.4017308	0.559347	0.72	0.4796	3.3648081
KLOC	0.0247329	0.003259	7.59	<.0001	1.5549672

Model #9: Cases with < 500 Person-months Effort, Project Descriptive Variables Only

Summary of Fit

RSquare	0.150433
RSquare Adj	-0.00886
Root Mean Square Error	138.07
Mean of Response	159.2154
Observations (or Sum Wgts)	39

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	108017.15	18002.9	0.9444
Error	32	610026.32	19063.3	Prob > F
C. Total	38	718043.47		0.4776

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	89.553561	72.05089	1.24	0.2229	.
DataCC	132.70347	75.48058	1.76	0.0883	1.7165455
DataP	-84.50281	69.93131	-1.21	0.2358	1.3024028
AvMon	38.633288	68.68684	0.56	0.5777	2.2844746
Simul	112.44482	78.5538	1.43	0.1620	1.4109775
MissDSP	55.174792	81.4375	0.68	0.5030	2.4085156
ComContrl	82.075885	69.00391	1.19	0.2430	1.5883125

Model #9 with Lines of Code Added

Summary of Fit

RSquare	0.875397
RSquare Adj	0.847261
Root Mean Square Error	53.72286
Mean of Response	159.2154
Observations (or Sum Wgts)	39

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	628572.97	89796.1	31.1128
Error	31	89470.50	2886.1	Prob > F
C. Total	38	718043.47		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	41.066591	28.26643	1.45	0.1563	.
DataCC	29.840889	30.35168	0.98	0.3331	1.8332885
DataP	-39.37025	27.41692	-1.44	0.1610	1.3222689
AvMon	-11.81527	26.98866	-0.44	0.6646	2.3296049
Simul	47.936965	30.94029	1.55	0.1315	1.4458227
MissDSP	-63.01488	32.8866	-1.92	0.0646	2.5942922
ComContrl	28.050518	27.14902	1.03	0.3095	1.623967
KLOC	3.4900684	0.259872	13.43	<.0001	1.1379039

Model #10: Cases with < 500 Person-months Effort, Project Descriptive Variables Only with Natural Log Transformation of Effort

Summary of Fit

RSquare	0.221422
RSquare Adj	0.075439
Root Mean Square Error	0.984888
Mean of Response	4.639606
Observations (or Sum Wgts)	39

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	8.827600	1.47127	1.5168
Error	32	31.040155	0.97000	Prob > F
C. Total	38	39.867755		0.2044

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	4.2241348	0.513957	8.22	<.0001	.
DataCC	1.1187554	0.538422	2.08	0.0458	1.7165455
DataP	-0.660567	0.498838	-1.32	0.1948	1.3024028
AvMon	0.0061626	0.489961	0.01	0.9900	2.2844746
Simul	0.9603011	0.560344	1.71	0.0962	1.4109775
MissDSP	0.2952795	0.580914	0.51	0.6147	2.4085156
ComContrl	0.5980022	0.492222	1.21	0.2333	1.5883125

Model #10 with Lines of Code Added

Summary of Fit

RSquare	0.756857
RSquare Adj	0.701954
Root Mean Square Error	0.559192
Mean of Response	4.639606
Observations (or Sum Wgts)	39

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	30.174192	4.31060	13.7853
Error	31	9.693563	0.31270	Prob > F
C. Total	38	39.867755		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	3.913639	0.29422	13.30	<.0001	.
DataCC	0.4600548	0.315925	1.46	0.1554	1.8332885
DataP	-0.371551	0.285378	-1.30	0.2025	1.3222689
AvMon	-0.316895	0.28092	-1.13	0.2680	2.3296049
Simul	0.5472125	0.322052	1.70	0.0993	1.4458227
MissDSP	-0.461571	0.342311	-1.35	0.1873	2.5942922
ComContrl	0.2520402	0.282589	0.89	0.3793	1.623967
KLOC	0.0223493	0.002705	8.26	<.0001	1.1379039

Model #11: Cases with < 500 Person-months Effort, COCOMO Factor-level Variables and EAF Only

Summary of Fit

RSquare	0.264058
RSquare Adj	-0.03577
Root Mean Square Error	139.8992
Mean of Response	159.2154
Observations (or Sum Wgts)	39

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	11	189605.25	17236.8	0.8807
Error	27	528438.22	19571.8	Prob > F
C. Total	38	718043.47		0.5690

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	424.17123	138.4096	3.06	0.0049	.
DATAL	-163.4845	104.5806	-1.56	0.1296	5.4162521
TIMEH	-5.647326	120.8546	-0.05	0.9631	2.6789181
STORH	260.66628	203.0787	1.28	0.2102	3.9982072
TURNH	-26.89614	112.5216	-0.24	0.8129	4.478584
ACAPL	-40.34196	110.4337	-0.37	0.7177	4.9210603
AEXPL	-46.9819	81.34575	-0.58	0.5683	3.1208809
PCAPVL	107.72333	118.4752	0.91	0.3713	2.5744717
LEXPN	-182.8063	111.4513	-1.64	0.1126	4.3937894
MODPH	-48.38109	141.6906	-0.34	0.7354	3.6822665
SCEDVH	-49.59474	82.03073	-0.60	0.5505	3.3499757
EAF	-96.88175	133.7734	-0.72	0.4752	3.2662729

Model #11 with Lines of Code Added

Summary of Fit

RSquare	0.90212
RSquare Adj	0.856945
Root Mean Square Error	51.9918
Mean of Response	159.2154
Observations (or Sum Wgts)	39

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	12	647761.65	53980.1	19.9694
Error	26	70281.82	2703.1	Prob > F
C. Total	38	718043.47		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	124.52284	56.35292	2.21	0.0361	.
DATAL	-40.34178	40.00051	-1.01	0.3225	5.7370494
TIMEH	9.9123065	44.93001	0.22	0.8271	2.680815
STORH	26.758102	77.58081	0.34	0.7329	4.224799
TURNH	55.478875	42.29325	1.31	0.2011	4.58112
ACAPL	-19.38098	41.07287	-0.47	0.6410	4.9286337
AEXPL	-25.10385	30.27781	-0.83	0.4146	3.1305245
PCAPVL	-102.069	46.88609	-2.18	0.0388	2.9193205
LEXPN	18.673403	44.21631	0.42	0.6763	5.0071953
MODPH	52.275501	53.22214	0.98	0.3350	3.7616507
SCEDVH	3.8181236	30.76053	0.12	0.9022	3.4106489
EAF	-86.2431	49.72193	-1.73	0.0947	3.2671554
KLOC	3.8429667	0.295185	13.02	<.0001	1.5675612

Model #12: Cases with < 500 Person-months Effort, COCOMO Factor-level Variables and EAF Only with Natural Log Transformation of Effort

Summary of Fit

RSquare	0.238616
RSquare Adj	-0.07158
Root Mean Square Error	1.060305
Mean of Response	4.639606
Observations (or Sum Wgts)	39

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	11	9.513086	0.86483	0.7692
Error	27	30.354669	1.12425	Prob > F
C. Total	38	39.867755		0.6664

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	5.6640957	1.049016	5.40	<.0001	.
DATAL	-1.094524	0.792623	-1.38	0.1786	5.4162521
TIMEH	0.2408075	0.915965	0.26	0.7946	2.6789181
STORH	1.0894831	1.539147	0.71	0.4851	3.9982072
TURNH	-0.167744	0.852809	-0.20	0.8455	4.478584
ACAPL	-0.218548	0.836984	-0.26	0.7960	4.9210603
AEXPL	-0.038169	0.616525	-0.06	0.9511	3.1208809
PCAPVL	1.1766178	0.897932	1.31	0.2011	2.5744717
LEXPN	-1.141272	0.844697	-1.35	0.1879	4.3937894
MODPH	-0.266768	1.073883	-0.25	0.8057	3.6822665
SCEDVH	-0.252337	0.621716	-0.41	0.6880	3.3499757
EAF	-0.111223	1.013877	-0.11	0.9135	3.2662729

Model #12 with Lines of Code Added

Summary of Fit

RSquare	0.785385
RSquare Adj	0.686332
Root Mean Square Error	0.57366
Mean of Response	4.639606
Observations (or Sum Wgts)	39

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	12	31.311529	2.60929	7.9289
Error	26	8.556226	0.32909	Prob > F
C. Total	38	39.867755		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	3.5972035	0.621779	5.79	<.0001	.
DATAL	-0.24512	0.441352	-0.56	0.5834	5.7370494
TIMEH	0.3481335	0.495743	0.70	0.4888	2.680815
STORH	-0.523951	0.856	-0.61	0.5458	4.224799
TURNH	0.4004563	0.466649	0.86	0.3987	4.58112
ACAPL	-0.073965	0.453184	-0.16	0.8716	4.9286337
AEXPL	0.1127393	0.334075	0.34	0.7385	3.1305245
PCAPVL	-0.270472	0.517325	-0.52	0.6055	2.9193205
LEXPN	0.2484799	0.487868	0.51	0.6148	5.0071953
MODPH	0.4275332	0.587235	0.73	0.4731	3.7616507
SCEDVH	0.1160907	0.339401	0.34	0.7351	3.4106489
EAF	-0.03784	0.548615	-0.07	0.9455	3.2671554
KLOC	0.0265077	0.003257	8.14	<.0001	1.5675612

Model #13: Cases with > 500 Person-months Effort, Project Descriptive Variables, COCOMO Factor-level Variables and EAF

Summary of Fit

RSquare	0.894236
RSquare Adj	0.153889
Root Mean Square Error	791.9596
Mean of Response	1527.933
Observations (or Sum Wgts)	9

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	5302993.5	757571	1.2079
Error	1	627200.0	627200	Prob > F
C. Total	8	5930193.5		0.6069

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	973	791.9596	1.23	0.4349	.
DataCC	275	1120	0.25	0.8467	3.1111111
DataP	-67	1120	-0.06	0.9620	3.1111111
Simul	-401.6	1120	-0.36	0.7808	4.4444444
MissDSP	1863	1481.621	1.26	0.4277	5.4444444
ComContrl	395	1120	0.35	0.7841	3.1111111
RELYL	932	1120	0.83	0.5582	1.7777778
TIMEH	245.6	1583.919	0.16	0.9021	8

Adding Lines of Code to Model #13 resulted in an infeasible model.

Model #14: Cases with > 500 Person-months Effort, Project Descriptive Variables, COCOMO Factor-level Variables and EAF with Natural Log Transformation of Effort

Summary of Fit

RSquare	0.963562
RSquare Adj	0.708498
Root Mean Square Error	0.299924
Mean of Response	7.195137
Observations (or Sum Wgts)	9

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	2.3787700	0.339824	3.7777
Error	1	0.0899547	0.089955	Prob > F
C. Total	8	2.4687247		0.3773

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	6.8803841	0.299924	22.94	0.0277	.
DataCC	0.2489135	0.424157	0.59	0.6622	3.1111111
DataP	-0.055181	0.424157	-0.13	0.9176	3.1111111
Simul	-0.532295	0.424157	-1.25	0.4283	4.4444444
MissDSP	1.195996	0.561107	2.13	0.2793	5.4444444
ComContrl	0.340721	0.424157	0.80	0.5692	3.1111111
RELYL	0.5195593	0.424157	1.22	0.4359	1.7777778
TIMEH	0.3271644	0.599849	0.55	0.6821	8

Adding Lines of Code to Model #14 resulted in an infeasible model.

Model #15: Cases with > 500 Person-months Effort, Project Descriptive Variables Only

Summary of Fit

RSquare	0.818456
RSquare Adj	0.515882
Root Mean Square Error	599.0525
Mean of Response	1527.933
Observations (or Sum Wgts)	9

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	4853601.7	970720	2.7050
Error	3	1076591.8	358864	Prob > F
C. Total	8	5930193.5		0.2211

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	973	599.0525	1.62	0.2028	.
DataCC	213.6	792.472	0.27	0.8050	2.7222222
DataP	55.8	599.0525	0.09	0.9317	1.5555556
Simul	-340.2	792.472	-0.43	0.6967	3.8888889
MissDSP	2047.2	669.7611	3.06	0.0551	1.9444444
ComContrl	861	733.6865	1.17	0.3253	2.3333333

Model #15 with Lines of Code Added

Summary of Fit

RSquare	0.991711
RSquare Adj	0.966843
Root Mean Square Error	156.776
Mean of Response	1527.933
Observations (or Sum Wgts)	9

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	5881036.1	980173	39.8789
Error	2	49157.4	24579	Prob > F
C. Total	8	5930193.5		0.0247

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-995.2194	342.42	-2.91	0.1008	.
DataCC	906.83216	233.4719	3.88	0.0603	3.4498157
DataP	-193.4248	161.4453	-1.20	0.3536	1.6495959
Simul	1133.8905	308.2122	3.68	0.0666	8.5887139
MissDSP	505.35903	295.962	1.71	0.2298	5.5436825
ComContrl	391.93192	205.2598	1.91	0.1964	2.6664552
KLOC	6.91331	1.069273	6.47	0.0231	4.8466807

Model #16: Cases with > 500 Person-months Effort, Project Descriptive Variables Only with Natural Log Transformation of Effort

Summary of Fit

RSquare	0.898051
RSquare Adj	0.728135
Root Mean Square Error	0.289646
Mean of Response	7.195137
Observations (or Sum Wgts)	9

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	2.2170399	0.443408	5.2853
Error	3	0.2516848	0.083895	Prob > F
C. Total	8	2.4687247		0.1006

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	6.8803841	0.289646	23.75	0.0002	.
DataCC	0.1671224	0.383166	0.44	0.6922	2.7222222
DataP	0.1084015	0.289646	0.37	0.7331	1.5555556
Simul	-0.450503	0.383166	-1.18	0.3245	3.8888889
MissDSP	1.4413694	0.323834	4.45	0.0211	1.9444444
ComContrl	0.6005007	0.354743	1.69	0.1891	2.3333333

Model #16 with Lines of Code Added

Summary of Fit

RSquare	0.981358
RSquare Adj	0.925434
Root Mean Square Error	0.151692
Mean of Response	7.195137
Observations (or Sum Wgts)	9

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	2.4227037	0.403784	17.5478
Error	2	0.0460210	0.023011	Prob > F
C. Total	8	2.4687247		0.0549

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	5.9997909	0.331316	18.11	0.0030	.
DataCC	0.4772786	0.225901	2.11	0.1690	3.4498157
DataP	-0.003103	0.15621	-0.02	0.9860	1.6495959
Simul	0.2090135	0.298218	0.70	0.5559	8.5887139
MissDSP	0.7515405	0.286365	2.62	0.1197	5.5436825
ComContrl	0.3906368	0.198604	1.97	0.1881	2.6664552
KLOC	0.0030931	0.001035	2.99	0.0960	4.8466807

Model #17: Cases with > 500 Person-months Effort, COCOMO Factor-level Variables and EAF Only

Summary of Fit

RSquare	0.894236
RSquare Adj	0.153889
Root Mean Square Error	791.9596
Mean of Response	1527.933
Observations (or Sum Wgts)	9

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	5302993.5	757571	1.2079
Error	1	627200.0	627200	Prob > F
C. Total	8	5930193.5		0.6069

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	1181	791.9596	1.49	0.3761	.
RELYL	3049	1481.621	2.06	0.2880	3.1111111
CPLXL	401.6	1120	0.36	0.7808	3.1111111
TIMEH	-498	1120	-0.44	0.7336	4
STORH	676.6	1120	0.60	0.6540	4.4444444
VIRTH	1715.4	1857.31	0.92	0.5253	4.8888889
TURNH	1320.4	1481.621	0.89	0.5366	7
ACAPL	-1930	969.9485	-1.99	0.2965	3

Adding Lines of Code to Model #17 resulted in an infeasible model.

Model #18: Cases with > 500 Person-months Effort, COCOMO Factor-level Variables and EAF Only with Natural Log Transformation of Effort

Summary of Fit

RSquare	0.963562
RSquare Adj	0.708498
Root Mean Square Error	0.299924
Mean of Response	7.195137
Observations (or Sum Wgts)	9

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	2.3787700	0.339824	3.7777
Error	1	0.0899547	0.089955	Prob > F
C. Total	8	2.4687247		0.3773

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	7.0741168	0.299924	23.59	0.0270	.
RELYL	1.9177244	0.561107	3.42	0.1812	3.1111111
CPLXL	0.5322946	0.424157	1.25	0.4283	3.1111111
TIMEH	-0.509224	0.424157	-1.20	0.4421	4
STORH	0.7812081	0.424157	1.84	0.3167	4.4444444
VIRTH	0.8658705	0.703385	1.23	0.4343	4.8888889
TURNH	0.5251495	0.561107	0.94	0.5211	7
ACAPL	-1.251177	0.367331	-3.41	0.1818	3

Adding Lines of Code to Model #18 resulted in an infeasible model.

Appendix D. ISBSG Model Regression Outputs

Model #1: All Cases

Summary of Fit

RSquare	0.512898
RSquare Adj	0.310564
Root Mean Square Error	6053.368
Mean of Response	4543.667
Observations (or Sum Wgts)	93

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	27	2507952123	92887116	2.5349
Error	65	2381811892	36643260	Prob > F
C. Total	92	4889764015		0.0012

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	2290.6003	1238.34	1.85	0.0689	.
Transaction/Production System	2008.9993	1850.083	1.09	0.2815	1.8280017
GUI Interface Application	13505.819	6612.66	2.04	0.0452	1.1804924
Sales	-6238.365	6758.954	-0.92	0.3594	1.233303
Research & Development	-3495.185	8743.83	-0.40	0.6907	2.0640215
Sales & Marketing	-8353.4	10557.62	-0.79	0.4317	3.009141
Accounting	24191.611	7509.678	3.22	0.0020	1.5224864
Engineering	9681.0216	7774.491	1.25	0.2175	7.8040419
Telecommunications	-9850.999	7641.184	-1.29	0.2019	4.6260258
Logistics	16151.623	11852.64	1.36	0.1777	3.7926335
Medical and Health Care	-1997.6	6178.733	-0.32	0.7475	1.0306464
Manufacturing	8402.6027	11221.97	0.75	0.4567	3.3997643
PC	-1366.419	2649.027	-0.52	0.6077	2.6604692
Process Modelling	-29.20301	4099.877	-0.01	0.9943	3.7289482
JAD (Joint Application Development)	-144.3145	2009.811	-0.07	0.9430	1.3868179
Structured	515.81862	6612.66	0.08	0.9381	1.1804924
New Development	3711.1847	1779.901	2.09	0.0410	1.7254121
COBOL	3920.8066	2435.333	1.61	0.1123	1.3157143
FOCUS	-2377.6	6308.231	-0.38	0.7075	2.1252473
PL/I	-3126.1	4633.804	-0.67	0.5023	1.1467526
NATURAL	43681.389	11590.64	3.77	0.0004	3.6268137
ACCESS	-10393.2	9491.865	-1.09	0.2776	7.1382188
UNIX SCRIPT	4828.4004	6308.231	0.77	0.4468	1.0743008
C	20487.419	12699.57	1.61	0.1115	4.3540008
APG	-77.39729	7255.759	-0.01	0.9915	2.8116421
Communication	6851.3997	6178.733	1.11	0.2716	3.9881535
Public Administration	-15191.79	7830.345	-1.94	0.0567	7.9165776
Electricity, Gas, Water	-3608.87	9086.42	-0.40	0.6925	2.2289299

Model #1 with Functional Points Added

Summary of Fit

RSquare	0.75729
RSquare Adj	0.651105
Root Mean Square Error	4306.235
Mean of Response	4543.667
Observations (or Sum Wgts)	93

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	28	3702970053	132248930	7.1318
Error	64	1186793962	18543656	Prob > F
C. Total	92	4889764015		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-739.854	958.4061	-0.77	0.4430	.
Transaction/Production System	2670.2672	1318.684	2.02	0.0470	1.8351626
GUI Interface Application	13166.35	4704.293	2.80	0.0068	1.1805878
Sales	-4404.399	4813.598	-0.91	0.3636	1.2360873
Research & Development	-10798.74	6286.355	-1.72	0.0907	2.1081785
Sales & Marketing	-7275.298	7511.66	-0.97	0.3364	3.0101032
Accounting	11913.863	5556.842	2.14	0.0358	1.647273
Engineering	5441.3536	5555.764	0.98	0.3311	7.8752057
Telecommunications	-7286.456	5445.152	-1.34	0.1856	4.6420039
Logistics	-2141.451	8734.212	-0.25	0.8071	4.0696489
Medical and Health Care	803.46298	4409.245	0.18	0.8560	1.0371414
Manufacturing	3780.6764	8003.8	0.47	0.6383	3.4174481
PC	-1109.659	1884.732	-0.59	0.5581	2.6612356
Process Modelling	2199.7192	2929.75	0.75	0.4555	3.7627434
JAD (Joint Application Development)	2128.7891	1457.506	1.46	0.1490	1.4412139
Structured	2494.8375	4710.558	0.53	0.5982	1.1837345
New Development	3575.9222	1266.295	2.82	0.0063	1.7257176
COBOL	4383.6683	1733.403	2.53	0.0139	1.3171716
FOCUS	-594.1795	4493.034	-0.13	0.8952	2.1304559
PL/I	-1469.664	3302.84	-0.44	0.6578	1.1512459
NATURAL	25554.87	8548.918	2.99	0.0040	3.8988076
ACCESS	-3871.392	6801.005	-0.57	0.5712	7.2415527
UNIX SCRIPT	5952.3214	4489.722	1.33	0.1896	1.0753465
C	16553.804	9047.475	1.83	0.0720	4.3668098
APG	-623.5373	5162.038	-0.12	0.9042	2.8121306
Communication	6043.6512	4396.569	1.37	0.1740	3.9902435
Public Administration	-11098.09	5593.631	-1.98	0.0515	7.9829255
Electricity, Gas, Water	-5484.55	6468.104	-0.85	0.3996	2.2318422
Function Points	4.0962676	0.510269	8.03	<.0001	1.4572563

Model #2: All Cases, Natural Log Transformation of Effort

Summary of Fit

RSquare	0.590246
RSquare Adj	0.420041
Root Mean Square Error	1.002414
Mean of Response	7.593301
Observations (or Sum Wgts)	93

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	27	94.08440	3.48461	3.4678
Error	65	65.31418	1.00483	Prob > F
C. Total	92	159.39859		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	7.3653079	0.205064	35.92	<.0001	.
Transaction/Production System	0.2132265	0.306366	0.70	0.4889	1.8280017
GUI Interface Application	2.2213963	1.095031	2.03	0.0466	1.1804924
Sales	-2.234802	1.119256	-2.00	0.0501	1.233303
Research & Development	-0.57795	1.447944	-0.40	0.6911	2.0640215
Sales & Marketing	-3.276193	1.7483	-1.87	0.0654	3.009141
Accounting	1.3283155	1.243573	1.07	0.2894	1.5224864
Engineering	1.626857	1.287425	1.26	0.2109	7.8040419
Telecommunications	-2.168767	1.26535	-1.71	0.0913	4.6260258
Logistics	5.6123679	1.96275	2.86	0.0057	3.7926335
Medical and Health Care	-1.685135	1.023174	-1.65	0.1044	1.0306464
Manufacturing	1.9644137	1.858314	1.06	0.2944	3.3997643
PC	-0.00964	0.438668	-0.02	0.9825	2.6604692
Process Modelling	0.2838172	0.678923	0.42	0.6773	3.7289482
JAD (Joint Application Development)	0.0123908	0.332817	0.04	0.9704	1.3868179
Structured	-0.08327	1.095031	-0.08	0.9396	1.1804924
New Development	0.6722602	0.294745	2.28	0.0258	1.7254121
COBOL	0.7869119	0.403282	1.95	0.0553	1.3157143
FOCUS	-0.017413	1.044618	-0.02	0.9868	2.1252473
PL/I	-0.542408	0.76734	-0.71	0.4822	1.1467526
NATURAL	7.7393435	1.919364	4.03	0.0001	3.6268137
ACCESS	-3.625939	1.571816	-2.31	0.0243	7.1382188
UNIX SCRIPT	1.5405675	1.044618	1.47	0.1451	1.0743008
C	3.7275961	2.102998	1.77	0.0810	4.3540008
APG	0.0397883	1.201525	0.03	0.9737	2.8116421
Communication	1.7553266	1.023174	1.72	0.0910	3.9881535
Public Administration	-5.204732	1.296674	-4.01	0.0002	7.9165776
Electricity, Gas, Water	-0.905439	1.504675	-0.60	0.5494	2.2289299

Model #2 with Functional Points Added

Summary of Fit

RSquare	0.729937
RSquare Adj	0.611784
Root Mean Square Error	0.820134
Mean of Response	7.593301
Observations (or Sum Wgts)	93

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	28	116.35088	4.15539	6.1779
Error	64	43.04771	0.67262	Prob > F
C. Total	92	159.39859		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	6.9516453	0.182531	38.08	<.0001	.
Transaction/Production System	0.3034908	0.251147	1.21	0.2313	1.8351626
GUI Interface Application	2.1750582	0.895946	2.43	0.0180	1.1805878
Sales	-1.984462	0.916763	-2.16	0.0342	1.2360873
Research & Development	-1.574899	1.197254	-1.32	0.1931	2.1081785
Sales & Marketing	-3.129031	1.430617	-2.19	0.0324	3.0101032
Accounting	-0.34762	1.058316	-0.33	0.7436	1.647273
Engineering	1.0481346	1.058111	0.99	0.3256	7.8752057
Telecommunications	-1.818702	1.037044	-1.75	0.0843	4.6420039
Logistics	3.1153296	1.663455	1.87	0.0657	4.0696489
Medical and Health Care	-1.302785	0.839753	-1.55	0.1257	1.0371414
Manufacturing	1.3335122	1.524346	0.87	0.3849	3.4174481
PC	0.0254086	0.358952	0.07	0.9438	2.6612356
Process Modelling	0.5880692	0.557979	1.05	0.2959	3.7627434
JAD (Joint Application Development)	0.3226736	0.277586	1.16	0.2494	1.4412139
Structured	0.1868698	0.897139	0.21	0.8357	1.1837345
New Development	0.6537967	0.24117	2.71	0.0086	1.7257176
COBOL	0.8500933	0.330131	2.58	0.0123	1.3171716
FOCUS	0.2260274	0.855711	0.26	0.7925	2.1304559
PL/I	-0.316301	0.629035	-0.50	0.6168	1.1512459
NATURAL	5.2650403	1.628165	3.23	0.0019	3.8988076
ACCESS	-2.7357	1.295271	-2.11	0.0386	7.2415527
UNIX SCRIPT	1.6939848	0.85508	1.98	0.0519	1.0753465
C	3.1906505	1.723117	1.85	0.0687	4.3668098
APG	-0.034761	0.983125	-0.04	0.9719	2.8121306
Communication	1.6450674	0.837339	1.96	0.0538	3.9902435
Public Administration	-4.645934	1.065323	-4.36	<.0001	7.9829255
Electricity, Gas, Water	-1.161473	1.231868	-0.94	0.3493	2.2318422
Function Points	0.0005591	0.000097	5.75	<.0001	1.4572563

Model #3: Cases with < Person-hours Effort

Summary of Fit

RSquare	0.476817
RSquare Adj	0.272864
Root Mean Square Error	1971.051
Mean of Response	2440
Observations (or Sum Wgts)	83

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	23	208903621	9082766	2.3379
Error	59	229217505	3885042	Prob > F
C. Total	82	438121126		0.0046

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	1812.5319	427.8802	4.24	<.0001	.
Transaction/Production System	94.282524	633.5001	0.15	0.8822	1.7623118
Sales	-4056.57	2209.819	-1.84	0.0714	1.2418053
Research & Development	-1050.083	2856.711	-0.37	0.7145	2.0752608
Sales & Marketing	-8831.468	3440.67	-2.57	0.0128	3.0104123
Engineering	-4890.214	1801.569	-2.71	0.0087	3.1806434
Telecommunications	-7936.283	2495.774	-3.18	0.0023	4.6360463
Logistics	1933.8601	2683.338	0.72	0.4739	1.8310108
Medical and Health Care	-1519.532	2016.959	-0.75	0.4542	1.0345087
Manufacturing	-3512.541	3235.027	-1.09	0.2820	2.6613113
PC	1289.6725	877.6382	1.47	0.1470	2.4364491
Process Modelling	1178.5698	1344.462	0.88	0.3843	2.5897937
JAD (Joint Application Development)	470.5193	668.0307	0.70	0.4840	1.4116243
Structured	-1662.204	2161.911	-0.77	0.4450	1.1885452
New Development	1266.0832	625.0714	2.03	0.0473	1.7569255
COBOL	900.15297	896.7486	1.00	0.3196	1.1521514
FOCUS	15.185528	2059.201	0.01	0.9941	2.1302891
PL/I	-733.3145	1515.845	-0.48	0.6303	1.1543854
ACCESS	2000.0093	2625.439	0.76	0.4492	5.1302819
UNIX SCRIPT	7221.1855	2059.201	3.51	0.0009	1.0782945
C	3647.9949	3272.487	1.11	0.2695	2.7233025
APG	-807.1017	2368.026	-0.34	0.7344	2.8171758
Communication	7329.4681	2016.959	3.63	0.0006	3.9866434
Electricity, Gas, Water	-1778.602	2976.406	-0.60	0.5524	2.2528085

Model #3 with Functional Points Added

Summary of Fit

RSquare	0.510594
RSquare Adj	0.308082
Root Mean Square Error	1922.727
Mean of Response	2440
Observations (or Sum Wgts)	83

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	24	223702166	9320924	2.5213
Error	58	214418960	3696879	Prob > F
C. Total	82	438121126		0.0021

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	1346.3065	478.0327	2.82	0.0066	.
Transaction/Production System	259.46872	623.4595	0.42	0.6788	1.7937685
Sales	-3510.008	2172.882	-1.62	0.1117	1.2617486
Research & Development	-2952.333	2944.404	-1.00	0.3202	2.3168357
Sales & Marketing	-8178.014	3372.169	-2.43	0.0184	3.038919
Engineering	-4736.207	1759.085	-2.69	0.0093	3.1867453
Telecommunications	-7222.359	2460.595	-2.94	0.0048	4.7356368
Logistics	-1579.971	3152.146	-0.50	0.6181	2.6552978
Medical and Health Care	-1115.821	1977.829	-0.56	0.5748	1.0453895
Manufacturing	-3765.949	3158.254	-1.19	0.2380	2.6655983
PC	1056.4055	864.0235	1.22	0.2264	2.4816357
Process Modelling	1476.9307	1319.951	1.12	0.2678	2.6232771
JAD (Joint Application Development)	731.42193	664.572	1.10	0.2756	1.4681518
Structured	-1179.28	2122.676	-0.56	0.5806	1.2041147
New Development	1141.0776	612.9392	1.86	0.0677	1.7753727
COBOL	1072.174	878.9782	1.22	0.2275	1.1632817
FOCUS	156.58957	2009.958	0.08	0.9382	2.1329262
PL/I	-626.5167	1479.644	-0.42	0.6736	1.1558898
ACCESS	2411.5354	2569.317	0.94	0.3518	5.1633728
UNIX SCRIPT	7182.8604	2008.807	3.58	0.0007	1.0783925
C	3708.3343	3192.398	1.16	0.2502	2.7235456
APG	-1006.51	2312.118	-0.44	0.6649	2.8224203
Communication	6749.6922	1988.735	3.39	0.0012	4.0731223
Electricity, Gas, Water	-1843.054	2903.612	-0.63	0.5281	2.2530858
Function Points	1.1163301	0.557957	2.00	0.0501	2.3281092

Model #4: Cases with < 10,000 Person-hours Effort, Natural Log Transformation of Effort

Summary of Fit

RSquare	0.587358
RSquare Adj	0.426498
Root Mean Square Error	0.829046
Mean of Response	7.317532
Observations (or Sum Wgts)	83

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	23	57.721669	2.50964	3.6514
Error	59	40.551737	0.68732	Prob > F
C. Total	82	98.273407		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	7.2095073	0.179971	40.06	<.0001	.
Transaction/Production System	0.1282961	0.266457	0.48	0.6320	1.7623118
Sales	-2.145454	0.929475	-2.31	0.0245	1.2418053
Research & Development	-0.418125	1.201565	-0.35	0.7291	2.0752608
Sales & Marketing	-3.431994	1.447184	-2.37	0.0210	3.0104123
Engineering	-3.642564	0.75776	-4.81	<.0001	3.1806434
Telecommunications	-2.083836	1.04975	-1.99	0.0518	4.6360463
Logistics	0.27384	1.128642	0.24	0.8091	1.8310108
Medical and Health Care	-1.529335	0.848356	-1.80	0.0765	1.0345087
Manufacturing	-2.993799	1.360688	-2.20	0.0317	2.6613113
PC	0.3015682	0.369144	0.82	0.4173	2.4364491
Process Modelling	0.5069617	0.565496	0.90	0.3736	2.5897937
JAD (Joint Application Development)	0.1742579	0.280981	0.62	0.5375	1.4116243
Structured	-0.238677	0.909324	-0.26	0.7939	1.1885452
New Development	0.5124353	0.262912	1.95	0.0560	1.7569255
COBOL	0.4981866	0.377183	1.32	0.1917	1.1521514
FOCUS	0.2233182	0.866123	0.26	0.7974	2.1302891
PL/I	-0.301677	0.637581	-0.47	0.6378	1.1543854
ACCESS	1.4880746	1.104289	1.35	0.1830	5.1302819
UNIX SCRIPT	1.7812985	0.866123	2.06	0.0442	1.0782945
C	-1.557648	1.376445	-1.13	0.2624	2.7233025
APG	-0.027556	0.996018	-0.03	0.9780	2.8171758
Communication	1.9111272	0.848356	2.25	0.0280	3.9866434
Electricity, Gas, Water	-0.907481	1.25191	-0.72	0.4714	2.2528085

Model #4 with Functional Points Added

Summary of Fit

RSquare	0.643503
RSquare Adj	0.495987
Root Mean Square Error	0.777199
Mean of Response	7.317532
Observations (or Sum Wgts)	83

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	24	63.239228	2.63497	4.3623
Error	58	35.034178	0.60404	Prob > F
C. Total	82	98.273407		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	6.9248253	0.193229	35.84	<.0001	.
Transaction/Production System	0.2291605	0.252013	0.91	0.3669	1.7937685
Sales	-1.811717	0.878315	-2.06	0.0436	1.2617486
Research & Development	-1.579658	1.190177	-1.33	0.1896	2.3168357
Sales & Marketing	-3.032989	1.363087	-2.23	0.0300	3.038919
Engineering	-3.548526	0.711052	-4.99	<.0001	3.1867453
Telecommunications	-1.647907	0.994614	-1.66	0.1030	4.7356368
Logistics	-1.871741	1.27415	-1.47	0.1472	2.6552978
Medical and Health Care	-1.282825	0.799472	-1.60	0.1140	1.0453895
Manufacturing	-3.148533	1.276619	-2.47	0.0166	2.6655983
PC	0.159133	0.349253	0.46	0.6504	2.4816357
Process Modelling	0.6891439	0.533546	1.29	0.2016	2.6232771
JAD (Joint Application Development)	0.3335677	0.268631	1.24	0.2193	1.4681518
Structured	0.0562014	0.858021	0.07	0.9480	1.2041147
New Development	0.4361056	0.24776	1.76	0.0836	1.7753727
COBOL	0.6032243	0.355298	1.70	0.0949	1.1632817
FOCUS	0.3096609	0.812459	0.38	0.7045	2.1329262
PL/I	-0.236465	0.598097	-0.40	0.6940	1.1558898
ACCESS	1.7393566	1.038561	1.67	0.0994	5.1633728
UNIX SCRIPT	1.7578968	0.811993	2.16	0.0345	1.0783925
C	-1.520804	1.290421	-1.18	0.2434	2.7235456
APG	-0.149316	0.934597	-0.16	0.8736	2.8224203
Communication	1.5571102	0.80388	1.94	0.0576	4.0731223
Electricity, Gas, Water	-0.946836	1.173689	-0.81	0.4231	2.2530858
Function Points	0.0006816	0.000226	3.02	0.0037	2.3281092

Model #5: Cases with > 10,000 Person-hours Effort

Summary of Fit

RSquare	0.780714
RSquare Adj	0.013214
Root Mean Square Error	10656.14
Mean of Response	22004.1
Observations (or Sum Wgts)	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	808559469	115508496	1.0172
Error	2	227106782	113553391	Prob > F
C. Total	9	1035666251		0.5795

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	14912.857	6976.083	2.14	0.1660	.
Transaction/Production System	19830.714	13952.17	1.42	0.2911	4.1142857
GUI Interface Application	-482.8571	12736.53	-0.04	0.9732	1.2857143
Accounting	32833.571	35112.18	0.94	0.4484	9.7714286
Engineering	-2353.286	20537.03	-0.11	0.9192	3.3428571
Process Modelling	-1983.571	18012.17	-0.11	0.9224	6
New Development	2576.4286	9865.672	0.26	0.8184	2.0571429
COBOL	-14254.29	13952.17	-1.02	0.4144	3.6

Model #5 with Functional Points Added

Summary of Fit

RSquare	0.96694
RSquare Adj	0.851229
Root Mean Square Error	4137.592
Mean of Response	22004.1
Observations (or Sum Wgts)	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	1001426913	143060988	8.3565
Error	2	34239338.1	17119669	Prob > F
C. Total	9	1035666251		0.1110

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-389.0679	4407.143	-0.09	0.9377	.
Transaction/Production System	12268.935	3138.512	3.91	0.0596	1.3809038
GUI Interface Application	12144.048	5695.233	2.13	0.1666	1.7051798
Engineering	-2975.871	6391.268	-0.47	0.6873	2.1474405
Process Modelling	9829.8555	3492.548	2.81	0.1065	1.4962653
New Development	12955.493	3250.359	3.99	0.0576	1.4810798
COBOL	-396.4073	3394.979	-0.12	0.9177	1.4138332
Function Points	3.5197633	0.852021	4.13	0.0539	1.6577435

Model #6: Cases with > 10,000 Person-hours Effort, Natural Log Transformation of Effort

Summary of Fit

RSquare	0.720271
RSquare Adj	-0.25878
Root Mean Square Error	0.582148
Mean of Response	9.88219
Observations (or Sum Wgts)	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	1.7452394	0.249320	0.7357
Error	2	0.6777926	0.338896	Prob > F
C. Total	9	2.4230321		0.6829

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	9.5495479	0.381105	25.06	0.0016	.
Transaction/Production System	0.874394	0.762211	1.15	0.3700	4.1142857
GUI Interface Application	0.0275167	0.6958	0.04	0.9720	1.2857143
Accounting	1.353493	1.918188	0.71	0.5535	9.7714286
Engineering	-0.256227	1.121944	-0.23	0.8406	3.3428571
Process Modelling	-0.026978	0.98401	-0.03	0.9806	6
New Development	0.1033427	0.538964	0.19	0.8656	2.0571429
COBOL	-0.542793	0.762211	-0.71	0.5503	3.6

Model #6 with Functional Points Added

Summary of Fit

RSquare	0.986699
RSquare Adj	0.940147
Root Mean Square Error	0.12694
Mean of Response	9.88219
Observations (or Sum Wgts)	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	2.3908043	0.341543	21.1956
Error	2	0.0322277	0.016114	Prob > F
C. Total	9	2.4230321		0.0458

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	8.7517037	0.13521	64.73	0.0002	.
Transaction/Production System	0.6001502	0.096289	6.23	0.0248	1.3809038
GUI Interface Application	0.6841374	0.174728	3.92	0.0595	1.7051798
Engineering	-0.129135	0.196083	-0.66	0.5778	2.1474405
Process Modelling	0.4267358	0.107151	3.98	0.0576	1.4962653
New Development	0.5863058	0.09972	5.88	0.0277	1.4810798
COBOL	0.0688428	0.104157	0.66	0.5766	1.4138332
Function Points	0.0001858	0.000026	7.11	0.0192	1.6577435

Bibliography

Angelis, L., I. Stamelos, and M. Morisio. "Building s Software Cost Estimation Model Based on Categorical Data," *Proceedings of the Seventh International Software Metrics Symposium*. IEEE Computer Society, 2001.

Barron, Gil. "Example COCOMO Applet." (20 August 1997). The COCOMO Model in a JAVA Applet. n.pag. 16 December 2005. http://www.cs.utexas.edu/users/lwerth/cs373/gil/example.html

Boehn, Marry W. *Software Engineering Economics*. Englewood Cliffs NJ: Prentice-Hall, Inc., 1981.

Ethiraj, Sendil K., Prashant Kale, M. S. Krishnan, and Jitendra V. Singh. "Determinants of Price in Custom Software: A Hedonic Analysis of Offshore Development Projects," *Social Science Research Network*, (June 2004). 29 July 2005 http://ssrn.com/abstract=569875

Foss, Tron, Erik Stensrud, Barbara Kitchenham, and Ingunn Myrtveit. "A Simulation Study of the Model Evaluation Criterion MMRE," *IEEE Transactions on Software Engineering*, 29: 985-995 (November 2003).

Jeffery, R., M. Ruhe, and I. Wieczorek. "A Comparative Study of Two Software Development Cost Modeling Techniques Using Multi-organizational and Company-Specific Data," *Information and Software Technology*, 42:1009-1016 (2000).

Jones, T. Capers. *Estimating Software Costs*. New York: McGraw-Hill, 1998.

Kemerer, Chris F. "An Empirical Validation of Software cost Estimation Models," *Communications of the ACM*, 30: 416-429 (May 1987).

Kennedy, Peter. *A Guide to Econometrics* (5th Edition). Cambridge MA: The MIT Press, 2003.

Lum, Karen T., and Jairus M. Hihn. "Economic Analysis of the Software Cost Function," *Space 2004 Conference and Exhibit*. September 2004.

-----. "The Structure of the Software Cost Function," *International Society of Parametric Analysis 2002 Conference Proceedings*. May 2002.

Makridakis, Spyros and others. *Forecasting Methods and Applications* (3rd Edition). Hoboken NJ: John Wiley and Sons, Inc., 1998.

McClave, James T. and others. *Statistics for Business and Economics* (9th Edition). Upper Saddle River NJ: Pearson Prentice Hall, 2005.

Parametric Estimating Handbook (2nd Edition). Spring 1999. 29 July 2005 http://www.ispa-cost.org/PEIWeb/newbook.htm

"Repository Data CD Release 8 – Field Descriptions." *R8 Data Disk Suite*. CD-ROM. International Software Benchmarking Standards Group Ltd., February 2003.

Sommerville, Ian. *Software Engineering* (7th Edition). Harlow England: Pearson Education Limited, 2004.

Stamelos, Ioannis, Lefteris Angelis, Maurizio Morisio, Evaggelos Sakellaris, and George L. Bleris. "Estimating the Development Cost of Custom Software," *Information and Management*, 40: 729-741 (2003).

"What you can find in the ISBSG Repository." *R8 Data Disk Suite*. CD-ROM. International Software Benchmarking Standards Group Ltd., February 2003.

Vita

Captain Marc Donovan Ellis was born in Leon, Iowa and grew up in Pierce, Nebraska. In 1997 he graduated for Pierce High School. He earned a Bachelor of Science degree in Management from the United States Air Force Academy in 2001 where he also obtained his commission in the United States Air Force.

His first assignment was to Sheppard AFB, Texas. While there, he served as a budget analyst and deputy Financial Analysis Officer and as the deputy Financial Services Officer. In the fall of 2004, he entered the cost analysis program of the Graduate School of Engineering and Management, Air Force Institute of Technology. Following graduation, Captain Ellis will be assigned to the Space and Missile Systems Center at Los Angeles AFB, California.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 074-0188*

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
23-03-2006	**Master's Thesis**	Oct 2004 – Mar 2006

4. TITLE AND SUBTITLE	**5a. CONTRACT NUMBER**
A Hedonic Approach to Estimating Software Cost Using Ordinary Least Squares Regression and Nominal Attribute Variables	**5b. GRANT NUMBER**
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	**5d. PROJECT NUMBER**
Ellis, Marc D., Capt, USAF	**5e. TASK NUMBER**
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)	**8. PERFORMING ORGANIZATION REPORT NUMBER**
Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 Hobson Way WPAFB OH 45433-7765	AFIT/GCA/ENV/06M-04
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**
NA	**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

12. DISTRIBUTION/AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Software spending is increasing within the DoD, NASA, and other technologically advanced organizations, with significant effects on program budgets. Cost estimators must have the best tools available. However, many current models are problematic due to inaccuracy and unavailability of the input parameters, the technical expertise and expense required to operate them, and the difficulty in explaining their outputs. Two databases were analyzed: 60 NASA/JPL software projects and 116 projects from the International Software Benchmarking and Standards Group database. Models developed using ordinary least squares regression employed parameters representing the presence of project characteristics. The models' predictive characteristics were compared based on the source database, project size, model transformations, and the variable combinations. COCOMO 81 estimates were calculated for comparison for the NASA/JPL projects. Few of the models met the mean absolute percentage error (MAPE) standard of 25 percent or less; however, managers may find mean error (ME) to be a better metric for evaluating software cost models. ME results for the NASA/JPL projects suggest that managers may prefer the simpler categorical models to the more complex models for smaller programs. The best of these models had 223.4 person-months, or 86.3 percent less error on average than the COCOMO based regression.

15. SUBJECT TERMS
Software Cost Estimating, Software Cost Models, Cost Analysis, Cost Models, Hedonic Analysis, Least Squares Regression, Regression Analysis, Development Effort, Software Engineering, Software Metrics

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Michael J. Hicks, Ph.D. (ENV)
REPORT	**ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(Include area code)*
U	U	U	UU	124	(937) 255-3636, ext 4605; e-mail: michael.hicks@afit.edu